Semantic and Structural Analysis of Web–based Learning Resources

Supporting Self–directed Resource–based Learning

Vom Fachbereich
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte

Dissertationsschrift

von

Dipl.–Inf. Philipp Claudius Friedrich–Eugen Scholl
Geboren am 21. Dezember 1979 in Stuttgart

Vorsitz: Prof. Dr.–Ing. Volker Hinrichsen
Erstreferent: Prof. Dr.–Ing. Ralf Steinmetz
Korreferent: Prof. Dr.–Ing. Wolfgang Effelsberg

Tag der Einreichung: 12. April 2011
Tag der Disputation: 17. Juni 2011

Darmstadt, 2011
Hochschulkennziffer D17

Abstract

In the *knowledge–based society*, the maintenance and acquisition of new knowledge are vital for each individual. Changed living and working conditions and the rapid development of technology cause the half–life of knowledge to decrease. Therefore, the knowledge that is acquired in educational institutions is no longer sufficient for an entire lifetime. Thus, self–directed learning at the workplace and in private life is becoming more and more important.

At the same time, the Web has become a very important source for knowledge acquisition, as it provides a huge amount of resources containing information that can be utilized for learning purposes. This form of self–directed learning that often involves learning with web resources is commonly referred to as *Resource–Based Learning*. In particular, it is characterized by a high degree of freedom in choice of resources and execution of the learning process.

When utilizing web resources as learning materials, learners face novel challenges: First, relevant information that covers the specific information need of a learner is often distributed over several web resources. This challenge can be addressed by providing adequate retrieval strategies where retrieval is not only restricted to a web search but also involves content that learners have already considered to be relevant. However, the so–called *vocabulary gap* — the fact that information can be expressed in completely different terminology, e.g. in technical terms or colloquial language — makes retrieval difficult. Further, in contrast to *Learning Objects* that are often used in educational institutions, web resources rarely include well–structured metadata. As Resource–Based Learning using web resources requires learners to handle and organize a large number of web resources efficiently, the availability of relevant metadata is vital. Eventually, in the majority of self–directed learning settings, the role of the teacher or tutor does not exist. These authorities usually set learning goals according to a curriculum, structure the learning process and assess the learning result. In self–directed learning, the learner has to take over these tasks which would otherwise have been accomplished by the teacher.

This thesis examines this form of Resource–Based Learning and derives adequate mechanisms to support this kind of learning. The requirements of supporting Resource–Based Learning are deduced and, based on these requirements, the design and the implementation of a tool called ELWMS.KOM is presented.

ELWMS.KOM is a tool that enables learners to organize their self–directed learning process and the contributing learning resources in a *personal knowledge network* by applying semantically *typed tags*. In particular, web resources are focused. Web resources are primarily not intended to be used for learning and thus, are rarely didactically adapted to learning scenarios. Further, they infrequently expose metadata that are relevant for learners. ELWMS.KOM is designed to attenuate these short–comings and the resulting challenges for learners by providing an appropriate level of support.

The contributions of this thesis comprise of the derivation and implementation of paradigms and technologies that enable such a supporting functionality in ELWMS.KOM. Based on an examination of Learning Objects that are commonly used in learning scenarios in educational institutions, the peculiarities and differences to self–directed learning paradigms are analysed and design decisions for ELWMS.KOM are inferred. These design decisions represent a foundation for the supporting functionalities that are proposed in this thesis.

Firstly, the technologies are presented that enable ELWMS.KOM to recommend tags and learning resources to the learner based on a semantic representation of their content. A user study based on ELWMS.KOM shows the need to support monolingual as well as cross–lingual approaches to recommend

semantically related tags and resources. An analysis of the approach that has been chosen to determine semantic relatedness is presented. Based on this analysis, several strategies are compared that show potential to reduce the computational complexity of this approach without considerably reducing its quality. Additionally, several extensions to improve the quality this approach that incorporate supplementary semantic properties of a reference corpus are presented and evaluated.

Furthermore, this thesis presents an approach to automatically segment web resources in order to support learners in the selection of relevant fragments of a web resource. This segmentation is based on a structural and visual analysis of web resources and yields a set of coherent segments. A user study confirms the quality of this approach.

In addition, an approach is introduced that supports learners in the consistent creation of their tagging vocabulary in ELWMS.KOM for the semantic tag type *Type*. This approach automatically recognizes the web genre of a web resource and is language–independent. Novel features have been developed that allow a reliable classification of web genres. Several evaluations using different feature sets and corpora are presented.

Finally, this thesis introduces the tag type *Goal* that supports learners to plan, execute and evaluate their overall learning process. This support feature has been derived from the theory of Self–Regulated Learning and has been implemented accordingly in ELWMS.KOM. The benefits are shown in two large–scale user studies that have been executed with ELWMS.KOM and the implemented goal setting mechanisms.

Zusammenfassung

In der Wissensgesellschaft wird die Pflege und Aneignung von Wissen für den Einzelnen immer wichtiger. Geänderte Lebens– und Arbeitsbedingungen sowie der technologische Fortschritt bedingen eine Abnahme der Halbwertszeit von Wissen. Somit genügt das in institutionellen Bildungseinrichtungen erworbene Wissen nicht mehr den sich ständig ändernden Bedingungen. Darum gewinnt selbstgesteuertes Lernen im Privaten oder am Arbeitsplatz immer mehr an Bedeutung.

Gleichzeitig wird insbesondere das Internet zu einer wichtigen Quelle von Lernmaterialien, weil es eine Vielzahl von Ressourcen umfasst, die potenziell zum Lernen eingesetzt werden können. Die Art von selbstgesteuertem Lernen, die unter anderem auf Webressourcen basiert, wird üblicherweise als *Ressourcenbasiertes Lernen* bezeichnet und ist durch einen hohen Freiheitsgrad in der Auswahl der Ressourcen und der Planung des Lernprozesses charakterisiert.

Mit der Nutzung von Webressourcen als Lernmaterialien stellen sich den Lernenden allerdings neue Herausforderungen: Erstens sind relevante Informationen, die den spezifischen Wissensbedarf eines Lernenden decken, oft über viele Webressourcen verteilt. Dies kann insbesondere durch eine Bereitstellung von geeigneten Suchmechanismen adressiert werden, wobei die Suche sich nicht auf eine Internetsuche beschränkt, sondern auch von Lernenden bereits gefundene und als relevant erachtete Ressourcen betrifft. Allerdings ist eine Suche oft durch die Nutzung unterschiedlicher Terminologie erschwert. Weiterhin sind Webressourcen (im Gegensatz zu oft in Bildungsinstitutionen eingesetzten Lernmaterialien in Form von *Lernobjekten*) meistens nicht durch wohlstrukturierte Metadaten beschrieben. Da Lernende mit einer Vielzahl von unterschiedlichen Ressourcen umgehen müssen, ist eine geeignete Beschreibung jedoch sehr wichtig, um eine angemessene Organisation der Lernressourcen zu erreichen. Zuletzt fehlt in selbstgesteuerten Lernszenarien meistens ein Lehrender oder Tutor, der Lernziele setzt, den Lernprozess strukturiert und das Lernergebnis bewertet. Der Lernende muss somit beim selbstgesteuerten Lernen die Aufgaben übernehmen, die ansonsten der Rolle des Lehrenden zufallen.

In dieser Arbeit wird diese Form des ressourcenbasierten Lernens betrachtet und geeignete Unterstützungsmöglichkeiten werden hierfür vorgestellt. Insbesondere werden aus den Eigenschaften des ressourcenbasierten Lernens die Anforderungen an ein Werkzeug zur Unterstützung, ELWMS.KOM genannt, herausgearbeitet und umgesetzt.

ELWMS.KOM ist ein System, das es Lernenden ermöglicht, ihren selbstgesteuerten Lernprozess und die dabei anfallenden Lernressourcen in einem *persönlichen Wissensnetz* mittels Auszeichnung der Ressourcen mit semantisch *typisierten Tags* zu organisieren. Insbesondere im Fokus stehen dabei webbasierte Ressourcen, die im Gegensatz zu Lernobjekten im bildungsinstitutionellen Kontext keine feste Struktur haben, nicht primär für Lernzwecke intendiert sind (und aus diesem Grunde nicht didaktisch aufbereitet sind) und nicht durch für den Lernenden wichtige Metadaten ausgezeichnet sind. ELWMS.KOM ist angelegt, diese Mängel und die daraus entstehenden Herausforderungen für den Lernenden durch angemessene Unterstützung abzumildern.

Die Beiträge dieser Arbeit umfassen die Herleitung und Umsetzung von Technologien und Paradigmen, die eine solche Unterstützung in ELWMS.KOM ermöglichen. Dazu werden, ausgehend von einer Analyse von Lernobjekten, die in bildungsinstitutionellen Lernszenarien verwendet werden, die Unterschiede zu freieren, selbstgesteuerten Lernparadigmen analysiert und auf dieser Basis Designentscheidungen für ELWMS.KOM abgeleitet. Diese bilden die Basis für die konkret in dieser Arbeit behandelten Unterstützungsmöglichkeiten.

Zum einen werden Technologien präsentiert, die es ELWMS.KOM erlauben, dem Lernenden Tags und Lernressourcen basierend auf einer semantischen Repräsentation ihres Inhalts vorzuschlagen. Dabei wird anhand einer Nutzerstudie die Notwendigkeit aufgezeigt, sowohl monolinguale als auch sprach–übergreifende Ansätze zur Ermittlung von semantisch ähnlichen Tags und Ressourcen zu ermöglichen. Eine Analyse des eingesetzten Ansatzes zur Ermittlung von semantischen Ähnlichkeiten wird präsentiert. Darauf aufbauend werden verschiedene Strategien vorgestellt und verglichen, die den Berechnungsaufwand der Methode reduzieren können, ohne die Güte des Ansatzes zu mindern. Weiterhin werden Erweiterungen für dieses Verfahren vorgestellt und evaluiert, die zusätzliche semantische Eigenschaften eines Referenzkorpus nutzen, um die Qualität des Ansatzes zu verbessern.

Ferner präsentiert diese Arbeit einen Ansatz zur automatischen Segmentierung von Webressourcen, um Lernenden die Auswahl relevanter Passagen einer Webressource zu vereinfachen. Diese Segmentierung baut auf einer strukturellen und visuellen Analyse von Webressourcen auf und hat eine Menge von kohärenten Segmenten zum Ergebnis. Eine Nutzerstudie belegt die Güte dieses Verfahrens.

Weiterhin unterstützt ein Ansatz Lernende bei der konsistenten Erstellung ihres in ELWMS.KOM verwendeten Tag–Vokabulars durch eine sprachunabhängige, automatisierte Erkennung des Web–Genres einer Webressource für den semantischen Tag–Typen *Typ*. Hier werden neuartige Merkmale entwickelt, anhand derer eine zuverlässige Klassifikation von Webressourcen in verschiedene Web–Genres möglich ist. Mehrere Evaluationen mit unterschiedlichen Parametrisierungen werden vorgestellt.

Zuletzt wird in dieser Arbeit der Tag–Typ *Ziel* eingeführt, der Lernende bei der Planung, Durchführung und Bewertung ihres gesamten Lernprozesses unterstützt. Diese Unterstützung in ELWMS.KOM wurde basierend auf der Lerntheorie des selbstregulierten Lernens hergeleitet und dementsprechend umgesetzt. Die Vorteile einer solchen Unterstützung werden in zwei großangelegten Studien nachgewiesen, die mit ELWMS.KOM und den darin integrierten Zielsetzungsmechanismen durchgeführt wurden.

Acknowledgements

This work would not have been possible without the support and encouragement of my advisors, colleagues, friends, family and students. I am very grateful to all those who contributed to the outcome of this thesis.

First of all, I would like to thank Prof. Ralf Steinmetz for his advice and for giving me the opportunity to do my research in this constructive and friendly atmosphere of the *Multimedia Communications Lab* (KOM). Furthermore, I thank Prof. Wolfgang Effelsberg for his valuable feedback as the second advisor of this thesis.

I am very grateful to Dr. Christoph Rensing for his support and knowledgeable advice over all the years, without you it would have taken me much longer. Further, I thank all friends and colleagues at the *Knowledge Media Group* for their support, the proof–reading of this thesis and generally for being part of a great team: Birgit, Jan, Lasse, Marek, Mojisola, Renato, Sebastian, Sonja, Stephan and Tomas.

I would particularly like to mention my colleague Doreen, who has been a steady companion over all the years. It was a great time and thank you for your constant encouragement! Further I thank Bastian with whom it was a real pleasure to do interdisciplinary research. I am very grateful to my colleagues at KOM (scientific as well as administrative) and at the *Research Training Group on Feedback Based Quality Management in e–Learning* for the companionship, the good times and the lively cooperation! I further want to thank my students for their contributions to my research and the many interesting and insightful discussions I had with you.

Last but not least, I am deeply indebted to my parents Margret and Stefan, my sister Onika and family, my brother Flori and all my friends who have supported me during the whole time. And, of course, a HUGE (!!!) thank you to Janine who always encouraged me to carry on and lifted the weight of everyday life during the last months, and Julius, who brought enjoyable distraction from work :)

Darmstadt, 2011 ϕl

Contents

1 Introduction

1.1 Motivation

In a knowledge–based society where knowledge is accumulating at a remarkable pace, the maintenance and acquisition of new knowledge are vital for each individual. Changed living and working conditions and the rapid development of technology cause the half–life of knowledge to decrease, and the knowledge that is acquired in educational institutions is no longer sufficient for an entire lifetime. Therefore, self–directed learning at the workplace and in private life is becoming more and more important. At the same time, the Web has become a very important source for knowledge acquisition, as it provides a huge amount of resources containing information that can be utilized for learning purposes. In [129], this form of self–directed learning using web resources is characterized as "the procurement of information and knowledge in order to solve current problems". Commonly, this type of learning is referred to as Resource–Based Learning (RBL). In its very nature, RBL is no new form of learning, as already learning involving textbooks or arbitrary digital or non–digital learning materials is subsumed by this notion. However, RBL often encompasses a high degree of freedom on the side of the learner as the learning process is executed in a *self–directed* way.

In this thesis, RBL is specified as learning that aims to match a learner's information need by self–directed interaction with a multitude of digital learning resources. Learning resources are defined as all digital resources that have the potential to support the process of learning. Especially digital resources that originate from the web preeminently are commonly used by learners, therefore in this thesis, RBL using web resources as learning materials constitutes the application scenario that is to be supported by the various approaches described in this thesis.

When using web resources, learners face novel challenges: First, relevant information that covers the information need of a learner is often dispersed over multiple web resources. As there is a huge amount of information available on the Web that is accumulating quickly, this can lead to information flooding. Providing an adequate retrieval in such a learning setting is therefore of upmost importance. Retrieval is not restricted to using information using web search engines, but also involves allowing learners to search in content they have already considered to be relevant. However, the so–called *vocabulary gap* — the fact that information can be expressed in completely different terminology, e.g. in technical terms or colloquial language — makes retrieval difficult, especially in cases when the learning materials consist of rather fine–granular web resource fragments.

Further, web resources do not have well–structured metadata like conventional Learning Objects that are stored in repositories. As RBL using web resources requires learners to handle a large number of web resources efficiently, the availability of relevant metadata is vital. Therefore, in RBL settings, learners should be enabled to characterize the resources with metadata in order to provide an adequate organization of resources.

Web resources are rarely didactically adapted to be used for learning, therefore they are neither structured nor specially geared towards learners as an audience. Further, the non–linear hyper–textual nature of the web can cause disorientation and add to the cognitive load of the learner. In order to accommodate to this challenge, the learner should be supported to structure and organize her learning adequately and functionality should be provided that helps the learner to proceed in a target–oriented way in her learning processes.

1.2 Goals and Contributions of this Thesis

The E–Learning KnoWledge Management System (ELWMS.KOM) is a platform supporting RBL using web resources that has been conceptualized and developed with the goal of assisting the different process steps that occur in such a self–directed learning setting. It addresses the above–mentioned challenges by providing learners with a tool set for structuring their learning process, executing the information search and enabling learners to adequately organize and persist their learning materials derived from web resources by the application of semantic tagging.

This thesis contributes to this support of self–directed RBL by addressing the above–mentioned challenges as follows:

- The conceptual design and implementation of the overall platform of ELWMS.KOM has been developed as a foundation for the support of RBL. Considerations including the nature and organization of learning materials used for RBL, didactic implications of self–direction and the roles of the learner have been incorporated into the design of ELWMS.KOM. Further, learning is often social in nature and paradigms like Computer–supported Collaborative Learning (CSCL) have emerged. ELWMS.KOM has been designed to support the community aspects of RBL, however, the focus of this thesis is on *personal* learning settings.
- Technology Enhanced Learning (TEL) in institutional learning settings is contrasted to self–directed learning. First, the notion of Learning Objects (LOs) is examined including implications on the use, roles and usage in TEL. Then, selected novel forms of learning that involve a changed role of the learner are presented and the paradigm of RBL is introduced. Based on an analysis of RBL, design objectives for ELWMS.KOM are derived and the implementation of ELWMS.KOM is briefly explicated.
- With long–term use, usually a large amount of web resources and tags composed in different languages accumulate, leading to disorientation and information flooding. This challenge can be met by recommending relevant web resources and tags in order to support the user to discover similar resources from other users or apply already–used tags. ELWMS.KOM already provides a basic *structural recommendation* approach, however, a *content–based recommendation approach* can additionally infer *implicit* connections between the resources and tags. In this thesis, the properties of the learning resources in ELWMS.KOM are analysed and on this basis, a generic approach is presented that is able to infer relations based on the semantic relatedness of web resources and tags. It is based on Explicit Semantic Analysis (ESA) [79] with the reference corpus of Wikipedia. This thesis systematically explores and evaluates the impact of concept and term reduction in ESA, reducing the overall computational complexity of the approach. Further, the applicability of ESA on cross–language setting is examined, providing a novel concept mapping approach and evaluating its performance. Eventually, ESA is enhanced by novel extensions to ESA that incorporate further semantic characteristics of Wikipedia. This approach named eXtended Explicit Semantic Analysis (XESA) is tested on a novel semantic corpus.
- As web resources often contain a lot of information that the learner does not necessarily require for meeting her information needs, ELWMS.KOM allows saving only the fragment of a web resource that is indeed relevant. For providing usability enhancements for the fragment selection process, an approach to automatically segmenting a web resource is presented. Related work is analysed and its short–comings are identified. Based on this, a novel approach to coherently segment web resources is presented, taking into account re–occurring structural patterns. An evaluation design is derived and in a user study the presented approach is evaluated.

- ELWMS.KOM supports the metadata assignment and organization of found web resources by semantic tagging. In this thesis, an approach to automatically detect the *web genre* of a web resource is presented that is able to distinguish between the genres *blog*, *wiki* and *forum*. An analysis of a social bookmarking application shows that the targeted genres belong to the most–used tags. Therefore, existing features are reviewed and novel features are presented that capture the pattern structure of a web resource. As the proposed approach exclusively takes into account structural features of the web resource, it proves to be language–independent. A corpus of multilingual instances of the targeted web genres is built and several evaluations are performed that support the applicability of this approach.
- ELWMS.KOM is targeted at supporting self–directed RBL. As in such a learning setting there is no didactic authority that plans or evaluates the learning process, this are tasks that the learner herself has to accomplish. Based on the theory of Self–Regulated Learning (SRL), the design of a tool to support self–directed learning is presented. Further, this thesis proposes a novel tag type for ELWMS.KOM that supports the learner to set her learning goals, namely the *Goal* type. The implementation of an extension to ELWMS.KOM is presented that augments ELWMS.KOM with so–called *scaffolds* that support learners to orchestrate their learning processes. Two large–scale user studies have been executed that show that supporting the processes of SRL is indeed beneficial for self–directed learners in RBL.

1.3 Structure of this Thesis

Chapter 2 presents the notion of RBL in more detail and explores the applicability of the concept of Learning Objects to novel learning paradigms. Further, it describes the design goals and the implementation of ELWMS.KOM, illustrating the links to the approaches presented in this thesis. Chapter 3 presents an application of semantic relatedness as a way of providing recommendations in ELWMS.KOM. In chapter 4, an algorithm to automatically segment web resources into coherent fragments is presented that enables usability support in ELWMS.KOM. Chapter 5 introduces a novel approach to web genre detection called Language–Independent Web Genre Detection (LIGD) that is language–independent and targets at supporting ELWMS.KOM to automatically classify the web genre of a resource in order to provide metadata. In chapter 6, an implementation of the notion of scaffolds in Self–Regulated Learning in ELWMS.KOM is introduced and its impact on the application of metacognitive processes is stated. Two user studies are presented that substantiate the theoretical benefits of supporting a goal–directed advancement in self–directed RBL. Chapter 7 gives a conclusion, revisits the contributions of this thesis and provides an outlook on future work.

2 Resource Based Learning and Learning Objects

This chapter discusses related work and states properties and requirements for supporting Resource–Based Learning (RBL) using web resources. Starting with e–learning in general, this section moves on to more specific topics concerning learning materials. Different definitions of Learning Objects (LOs) are compared, as well as content models and role allocation of the different stake–holders of the LO creation, maintaining and dissemination. Then, the implications on the codification, dissemination and handling of LOs that arise from a transition from learning in educational institutions towards personal learning are discussed. Based on this discussion, the notion of "Learning Resources" is introduced and reformulated in order to specify supporting functionality for Technology Enhanced Learning (TEL). Eventually, an overview of a system named E–Learning KnoWledge Management System (ELWMS.KOM) [176, 125] is given that has been designed to support the scenario that is targeted in this thesis, which is self–directed RBL with web resources. ELWMS.KOM addresses the challenges presented in chapter 1 by providing learners with a tool set for structuring their learning process, executing the information search and enabling learners to adequately organize and persist their learning materials derived from web resources by the application of semantic tagging.

2.1 Introduction

The applicability of computers (and related technologies) for educational purposes has been recognized from early on. Especially with the availability of personal computers, *e–learning* has increasingly crossed the border between learning in educational institutions and learning in the work or personal life. The term e–learning (or electronic learning) denotes any type of computer–supported learning, embracing computer–supported face–to–face learning as well as distance learning (e.g. with Computer Based Trainings (CBTs)), CSCL, formal and informal learning and various other pedagogical approaches and technologies. Another term that is often used interchangeably with e–learning is Technology Enhanced Learning (TEL), which is preferred in this thesis, as it emphasizes the *technological* aspects in the field of e–learning.

2.1.1 Settings of Technology Enhanced Learning

In the 1990s, CBTs often took advantage of reasonably fast computers that were able to display multimedia content like movies, pictures and simulation programs. As large bandwidth was not yet readily available, CBTs were usually shipped on electronic data storage media like CD–ROMs. Further, often these learning materials were compiled into self–contained data formats that were usually specific to the technologies the manufacturers used and supported.

With the advent of the Internet and broadband access to the "information super highway", things have quickly changed. Educational institutions are increasingly embracing the World Wide Web (WWW) as a means to make their learning materials available in an immediate, low–cost and accessible way, moving parts of their educational materials online as Web Based Training (WBT). Therefore, WBTs are usually created by using the (quasi–)standards of the web, e.g. HyperText Markup Language (HTML) [154], Cascading Stylesheets (CSS), Flash or ShockWave for multimedia content. Especially HTML provides a comfortable way to represent these learning materials, as it is an open and human–readable content presentation format that can be easily edited and provides the hyperlink structure in the Web.

TEL that is located in an institutional setting and leads to certification usually follows a specifically designed curriculum. The institution provides the learning materials, typically made available as WBTs in a Learning Management System (LMS), and structures the didactic approach entailing a systematic teaching and learning process.

However, changed living and working conditions and the outstanding development of technology cause the half–life of knowledge to decrease, and the knowledge that is acquired in educational institutions is not sufficient for a whole lifetime. Therefore, the importance of self–directed learning at the workplace or in personal learning settings is increasing enormously. Self–directed learners do not necessarily utilize (or do not have access to) dedicated learning resources and learning systems that have been explicitly designed for learning purposes, so they usually use the Web's vast amount of resources to search for information. This form of learning, called Resource–Based Learning, is often based on a personal information need and is executed by the learner autonomously.

These both learning settings pose different demands on the learning process. For one, RBL using web resources is multifaceted and does not follow an explicit curriculum or structure, therefore the support of its processes in learning applications is sparse. Further, this also affects the way learning materials are codified and disseminated. Due to the central importance of learning materials in RBL, these two points are examined in detail in the following sections.

2.1.2 Structure of this Chapter

In this chapter, the question is examined how this self–directed learning compares to the learning settings that have been predominantly targeted in research of TEL so far. Especially the differences of the learning materials' properties and the learning process itself in TEL are explored, because they are central to RBL in the addressed scenario. Section 2.2 introduces different facets of the notion of Learning Objects and lists a selection of the extensive related work in this field. In section 2.3, current directions of research are presented that do not see the learner in the role of a mere consumer of learning materials. Section 2.4 introduces the learning style Resource–Based Learning that encompasses the self–directed way of learning with web resources. In section 2.5, a novel tool called ELWMS.KOM aimed at supporting RBL in web–based learning settings is presented. Design decisions for its implementation are stated and a short overview of ELWMS.KOM's functionality is given. This chapter closes with a short conclusion in section 2.6.

2.2 Learning Objects

The definition of what constitutes a LO strongly depends on the usage scenario that is targeted by the respective researchers. Thus, as Polsani [150] states, there are multiple definitions of LOs that are not necessarily consistent and partly contradict each other.

These definitions of LOs cover different aspects of requirements that LOs have to fulfill. Often, granularity properties of LOs are mentioned, as well as the didactic dimension that focuses the intended usage of LOs in educational settings. Other authors highlight structural or functional properties of LOs (especially reusability) that they deem important.

2.2.1 Granularity

The broadest definition of LOs is given by the Learning Technology Standards Consortium (LTSC) of the Institute of Electrical and Electronics Engineers (IEEE) in its Learning Object Metadata (LOM) specifica-

tion [91]: a LO is "any entity, digital or non–digital, that may be used for learning, education or training". In version 4, LOM names "multimedia content, instructional content, instructional software [...and], in a wider sense, Learning Objectives, persons, organizations, or events" as examples for LOs. Wiley [196] advocates strongly that this definition is too broad to be of any practical or scientific value. He proposes a narrower definition, stating that a LO is "any *digital*[1] resource that can be reused to support learning", ignoring non–digital entities. Wiley gives examples for two granularity levels of LOs: smaller (e.g. images, photos, data feeds, video, small bits of text and programs like Java applets) and larger reusable digital resources (e.g. entire web pages that combine these smaller LOs in order to deliver a complete instructional event).

Polsani [150], however, denounces these granularity levels introduced by Wiley, classifying the smaller reusable digital resources rather as *digital assets* than LOs. A major critique is that Wiley's definition "appears to be a simple case of uncritical nomenclature" without following any conceptual direction. By classifying every digital asset as a LO, this would nullify the aspects of modularity, separation of content and context, and reusability.

Similarly, Boyle [35] argues that LOs have to consist of *fragments* that are topically *coherent*, meaning that each fragment should only have exactly one objective confined to a narrow scope, considering a single concept. Thus, according to this principle, LOs consist of multiple fragments that are subsequently aggregated. In practice, however, heavily fragmented LOs pose additional effort to organize and manage this multitude of fragments and open new challenges in the fields of LO retrieval and reuse [92].

2.2.2 Reusability

LOs have also a functional dimension. Especially reusability is a major concern for most researchers, as it allows a LO to be created once and being used in different contexts. For example, Polsani [150] puts a strong focus on reusability of LOs and states that there is a broad consensus in the understanding of functional requirements of LOs:

Reusability Once created, a LO should be reusable in different instructional contexts and settings.

Accessibility Metadata that describe and reference a LO should be added, so that it can be stored and referenced in a database.

Interoperability The LO should be independent of both the delivery media and LMS.

Further, Meyer [130] considers the scenario of reuse in which existing learning resources serve as preliminary products for the creation of new learning resources for WBTs. Specifically, he focuses on the multi–granularity reusability (i.e. the ability to separately reuse *parts* of a learning resource's content) of learning resources without restrictions to a particular authoring tool's format. This requires that parts of a learning resource with the potential to be reused remain available and retrievable for reuse (and therefore accessible), and that they are interoperable so that they can be aggregated to new learning resources.

2.2.3 Learning intention

L'Allier [112] defines a LO as the smallest, independent structural learning material that contains an objective (a statement of the result of a learning activity), a learning activity (the content needed to achieve the objective) and an assessment (a structural element of the LO that determines whether an objective

[1] Emphasis by the author.

has been met). Thus, according to L'Allier, a LO is always created with an educational intention, and therefore mirrors common properties of educational institutions: there is a curriculum (the objective), the learning content (the learning activity) and the assessment.

A less strict definition of Polsani [150] states that any digital object needs to be embedded in a *learning intention* in order to qualify as a LO. Although the term learning intention does not denote whether the intention has to be on the side of the teacher or the learner, it reflects L'Allier's notion of "objectives", but without any special instructional methodology. Further, Polsani states that creation and deployment of LOs should be exclusive processes, so that the LOs do not favour any instructional methodology and can be reused in multiple instructional contexts. A similar argumentation is presented by Baumgartner [18], who postulates that a LO has to be motivated by pedagogics and didactics and should at least communicate a specified *learning goal*. According to Baumgartner, only if these criteria are met, informational entities structured by content may be called "Learning Objects".

Koper [109], however, denounces the notion of LOs containing learning activities, objectives or prerequisites, as these hinder reusability in multiple contexts. He rather defines a LO as "any digital, reproducible and addressable resource used to *perform*[2] learning activities or learning support activities, made available for others to use", thus strictly separating the learning activities and the content.

2.2.4 Metadata for Learning Objects

Masie [126] highlights the need for assigning metadata to LOs. Metadata provides "the ability to richly describe and identify learning content so that we can find, assemble, and deliver the right learning content to the right person at the right time". Further, he defines LOs as reusable, media–independent chunks of information used as modular building blocks for e–learning content and states that LOs are most effective when organized by a metadata classification system and stored in a data repository such as a LMS. The Cisco content model presented by Barritt et al. [15] defines a minimal set of metadata that is needed for reusable LOs, namely a title, a level objective, specification of the major topic area, a job function or task and creation data like the date and the author.

As a joint effort by the LTSC of the IEEE and IMS Global Learning Consortium, the LOM Standard [91] satisfies the need of LOs to be described by metadata for enabling exchange and reusability over organizational borders and defines a set of commonly used metadata information that can be applied to nine categories:

General contains information about the LO as a whole, e.g. the topic or language of the LO.

Lifecycle groups the features related to history and current state of the LO, e.g. the version, creation date, date of last edit, authors and all other contributing persons.

Meta–Metadata describes the metadata itself, e.g. the version of the applied metadata scheme.

Technical contains information about technical characteristics of the LO, e.g. the format expressed by a Multipurpose Internet Mail Extensions (MIME) Type[3], its size in bytes or technical requirements (e.g. special applications that are needed to open a LO).

Educational groups educational and pedagogical characteristics for teachers, authors or learners, e.g. interactivity type of the LO or typical age range of the intended audience.

Rights describes intellectual property rights that the LOs may underlie, e.g. the costs and the copyright licence of the contents.

[2] Emphasis by the author.

[3] A MIME Type designates the content type of different file specifications, defined in the IETF's RFCs 2045–2047, 4288–4289 and 2049.

Relation contains features that specify the relation of a LO to other LOs, e.g. references to other documents.

Annotation provides annotations for LOs, e.g. an author's comments on the educational use of the LO.

Classification describes a LO in relation to a particular categorization system.

Hodgins [90] additionally differentiates between objective metadata and subjective metadata. Whereas most LOM categories describe (with the notable exception of the *Annotation* category) mainly objective metadata — i.e. metadata that can be derived from the LO itself or its intended usage —, Hodgins stresses the value of subjective metadata, i.e. metadata that codifies attributes that are subjective and are often determined by the person who creates this metadata. For example, subjective metadata may capture tacit knowledge, context, perspectives and opinions of a LO, that learners and teachers alike could profit from. Further, Hodgins postulates to "connect everything to everything" by usage of metadata, stating that LOs have an enormous potential to foster digital connectivity between LOs themselves and LOs and people. He theorizes what would be possible if control of content was put into the hands of every individual, if everyone in need of a given skill or knowledge could be connected directly with those who have it. "What will it mean to have potentially billions of authors and publishers?"[4]

2.2.5 Content Models

Commonly, the definition of LOs is implicated by a *content model*. Content models define the structure and the relations between different granularities of content. They define different kinds of LOs at different levels of granularity and base on the assumption that independent and self–contained learning content can be created. Thus, these LOs may be used alone or be dynamically assembled. Further, they can be combined to form longer educational interactions or even be reused in different learning contexts [193].

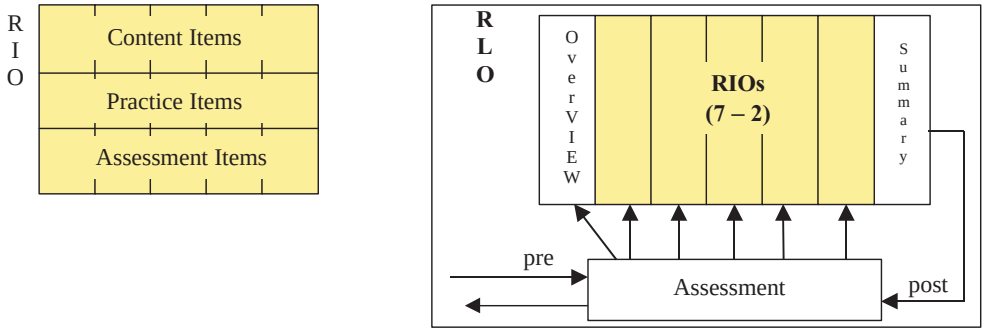


Figure 2.1: The Cisco Content Model with Reusable Information Objects enclosed in a Reusable Learning Object, cf. [15].

A content model that is frequently referenced is the Cisco content model [15]. Its purpose is to allow reusability by defining a content structure having fine granular Reusable Information Objects (RIOs) that can be aggregated into Reusable Learning Objects (RLOs). A RIO is classified as one of the classes *Concept*, *Fact*, *Process*, *Principle* or *Procedure*. A RLO is based on a single objective, derived from a specific job task. Each RIO is built upon an objective that supports the RLO's objective and contains content items, practice items and assessment items. RLOs wrap five to nine RIOs, adding an overview, a summary and assessments (see figure 2.1).

[4] Notably, Hodgins writes this before the notion of *Web 2.0* was coined.

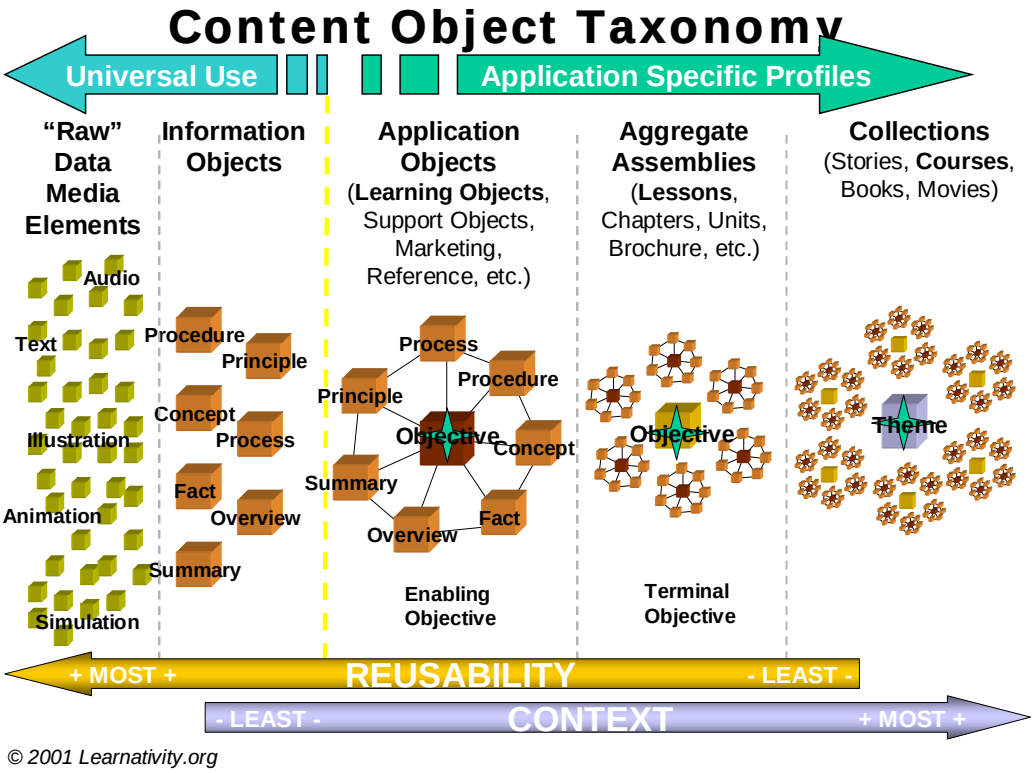


Figure 2.2: Autodesk Inc.'s Learnativity Component-based Content Model with the aggregation level *Learning Objects* as the central level, cf. [90].

Further, Hodgins [90] proposes the Learnativity Component–based Content Model that serves as a reference to Autodesk Inc.'s corporate content strategy. This model defines a five–level content hierarchy as shown in figure 2.2. It aims to be generic because it is not only intended to be used for learning materials, but it is also has been designed for representing marketing and reference materials as well. The content hierarchies describe different granularity and aggregation levels of the information materials and their chunks. The second level of *Information Objects* is formed by a set of these single information entities to create a granular, reusable chunk of information that is media independent. The third level, LOs (or more general *Application Objects*), aggregates Information Objects into meaningful objectives that already serve a certain didactic purpose.

In order to bridge the diversity of content model specifications, Verbert and Duval present ALOCOM [193], a generic content model for LOs addressing interoperability between different content model specifications. Based on comparative analysis of other content models, ALOCOM defines a generic model that maps these models via a generic ontology and allows sharing and reusing LOs on a global scale.

2.2.6 Learning Object Lifecycle

In the process of the creation, consumption and reuse of LOs, there are certain process steps that frequently occur. Rensing et al. [159] describe a lifecycle of LOs (see figure 2.3) that encompasses the steps of authoring, re–authoring, provision and learning in a systemic perspective.

This lifecycle model shows clearly the distinction between the author's and the learner's role as already described by Downes [62]: the learner is a mere consumer of learning materials and does rarely contribute to the creation or content maintenance of the LOs, as LMSs rarely implement feedback mech-

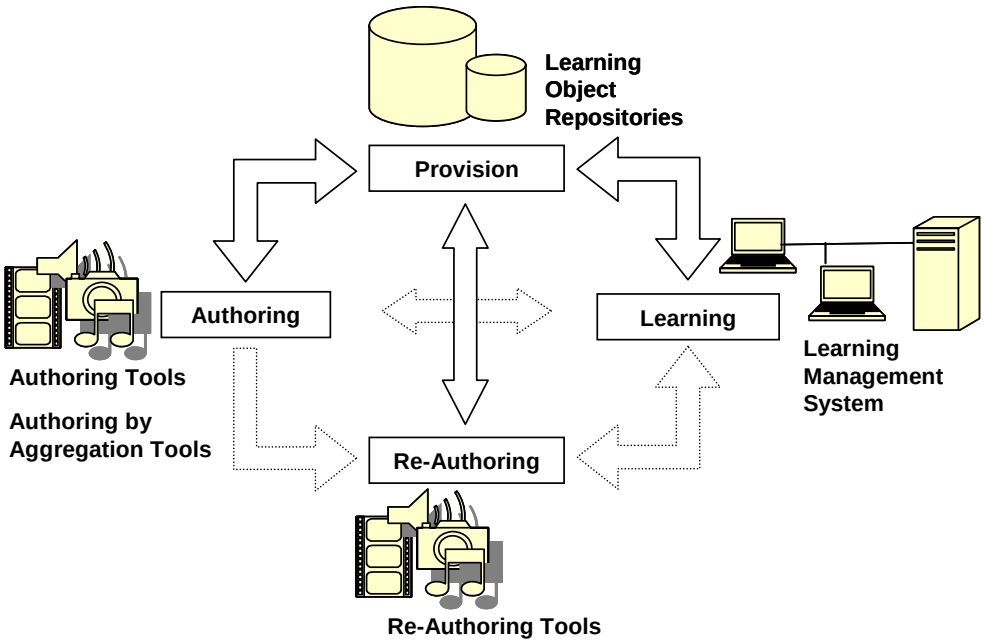


Figure 2.3: A lifecycle of Learning Objects, including authoring and reauthoring processes, showing the relation between creation and consumption of Learning Objects in a systemic perspective, cf. [159].

anisms that provide the learner with the possibility to communicate her questions to the author of a LO.

Traditionally in institutional learning settings, there is a sharp separation between author of a LO and the learner. This role allocation is typical for institutional learning settings but novel learning paradigms have emerged that aim at diffusing this differentiation.

2.3 The Emergence of Learner Participation

The related work presented in section 2.2 is mainly anchored in instructional settings that are based on different forms of institutional learning, i.e. commonly, there is an actor (e.g. a teacher, tutor or author of learning materials) who determines — based on pedagogical and didactic principles — what will be presented as learning materials and how learning performance will be assessed. In the last years[5], however, forms of learning have emerged that tend to put the organization and responsibility for the learning process into the hand of the learners. Although already institutional learning settings include episodes where learners have to learn autonomously (e.g. individually, when writing a thesis or seminar paper or collaboratively, in CSCL settings), the degree of individual responsibility for the own learning process is higher in non–institutional learning settings.

For example, in so–called *e–learning 2.0*, Downes [63] characterizes learning not only by greater autonomy for the learner, but also puts an emphasis on *active* learning, with creation, communication and participation playing key roles. Further, Downes states that the role of the teacher is changing, with the extreme of a collapse of the distinction between teacher and student altogether.

In this section, two of these novel learning forms or paradigms are presented and their impact on the nature of LOs is presented.

[5] Incidentally, this trend correlated with the emergence of the so–called *Web 2.0* that focuses on the emancipation of users by lowering the threshold to participate in the shaping of the communities in the Web, cf. [146].

2.3.1 Microlearning and Microcontents

Hug and Friesen [94] propose the term *microlearning* as a practice that is often encountered with learners using the Web in informal learning scenarios[6]. They refer to microlearning "... in terms of special moments or episodes of learning while dealing with specific tasks or content, and engaging in small but conscious steps". Microlearning does not represent a new conceptualization of learning, but rather targets the aspect of granularity of the learning episodes by a content model. It describes — in contrast to *meso–* and *macrolearning* (see table 2.1 for examples) — the way of learning by consuming fine–granular learning materials (so–called *microcontent*) in a relatively short learning episode. Microcontents often consist of chunks of content from Web 2.0 applications and social software, like blog posts and wiki pages [40] and deal with small or very small units and rather narrow topics of learning materials, having characteristics of fragments, facets, episodes, skill elements or discrete tasks [95]. Further, they are often addressable by an Uniform Resource Locator (URL) and convey one primary idea or concept. Learning using microcontent enables the learner to perform short and atomic learning activities, ensuring short feedback loops and allowing the learner an immediate and direct control over her learning process. Thus, microlearning rather aims to complement than to replace other learning conceptualizations.

	Linguistics	Language learning	Learning contents	Course Structure	Competency classification
micro level	single letters	vocabulary, phrases, sentences	learning objects, microcontent	learning objects	competencies of learners / teachers
meso level	words, letter–figure combinations, sentences	situations, episodes	sub areas, narrow topics	topics, lessons	designing a lecture
macro level	texts, conversations, linguistic communication	socio–cultural specifics, complex semantics	topics, subjects	courses, curricular structures	designing a curriculum

Table 2.1: Examples of microcontent for Microlearning — Mesolearning — Macrolearning, cf. [94].

How this microcontent is employed in a learning scenario strongly depends on the didactic design of a learning episode. Hug and Friesen [94] propose several models that conceptualize the didactics by positioning, combining, contextualizing and contrasting microcontent. These models can be seen in line with the content models presented in section 2.2. In the following, three selected models are listed:

- The *multicomponent model* is a — more or less systematically — organized combination of microcontent. Microcontents are forming linear or branching sequences, thus imposing a relation between components. For example, a *learning path* could be provided by a teacher for learners to navigate the microcontent.
- The *aggregation model* bundles microcontents that are inherently similar or related as a relatively unstructured entity. For example, a collection of microcontent could be assembled that cover a common topic.
- In the *emergence model*, coherent structures arise from relations between the microcontents themselves, forming novel patterns that originate from "a multiplicity of relatively simple interactions" [94]. For example, learners could explicitly create relations between the microcontents or different microcontents could be associated due to a common usage context. Eventually, the aggregated

[6] Microlearning is not exclusively employed in informal learning settings, yet it is often mentioned in combination with Web 2.0 applications, social software and wireless network technologies for consumption, cf. for example [40].

relations in their collectivity could evolve into a structure that may be used by other learners as well[7].

With regard to learner participation, microlearning does not stipulate a certain learning paradigm. However, as the emergence model suggests, informal and self–directed learning scenarios are supportable [94].

2.3.2 Personal Learning Environments

The notion of Personal Learning Environments (PLEs) has been conceptualized by Attwell [4] as a personal assembly of different applications, services and learning resources. This means that a learner autonomously selects the relevant *learning content*, the *educational providers* and the *context of learning* based on her own learning needs and tasks. Despite the conceptual imprecision of the notion of a PLE, the technology used is closely interwoven with applications and services that can be subsumed by the term *Web 2.0* [146] that has its root in the spirit of emancipating the user. The concept of PLEs builds strongly on the learning mode of Self–Directed Learning (SDL) [104], a strong autonomy of the learner and the possibility to personalize the PLE specifically to the need of the learner [192]. Attwell states that learning continues after having passed through the different educational institutions and has to take place every day over a whole life. Thus, learning cannot solely rely on the tools and infrastructure of one educational provider, and therefore the learner has to develop and organize her own learning processes, learning communities and learning styles. Further, PLEs aim at supporting learners specifically in SDL, as they do not entail certain systems or applications of educational institutions on the learner but are composed by different building blocks that together can form a learning environment.

Therefore, in PLEs, different learning styles are possible, namely "learning by personal interest or the desire to solve a problem, community learning, school learning, experiential learning, workplace learning, etc. In short, it can embrace all formal and informal learning" [87]. Due to their social aspect (most researchers see social software, online communities and connectivity between learners as a major driving force), PLEs have the potential to bridge personal and collective learning. This manifests in the choice of applications that are typically incorporated into a PLE, namely blogs, microblogs and wikis [40]. These applications allow content production, aggregation via *Really Simple Syndication (RSS)*–feeds and feedback via comments and trackbacks, enabling learners to participate in learning communities or communities of practice [115], e.g. in the *blogosphere* (the totality of all blogs, their interconnections and communities).

Thus, a PLE serves a learner not only as a "container" of learning content, but the learner herself is seen as a producer of learning contents alike, thus obliterating the borders between authors respectively consumers of learning materials. This is clearly in contrast to the strict separation between the provider or publisher of educational materials and the passive learner as seen in "classical" LOs as described by Downes [62]. As any type of content may be used for learning, there is no common content model, but Harmelen [192] describes a reference model that encompasses use cases and their infrastructure implications, services and software that may be employed and patterns that show how learning materials are incorporated into a PLE.

The personal, community–based nature of PLEs entails the primary usage of web resources as learning materials. As the learners become producers and mostly use the common content production systems of the web, a large part of the learning materials are again web–based. PLEs as a paradigm do not establish

[7] This is an effect that is, in general, postulated as a result (online) learning communities, where the learners will benefit from other learners' interactions and outcomes (cf. [4]).

a certain way of persisting or representing these learning materials, however, most implementations of PLEs (like e.g. Elgg[8]) use the common publishing functionalities like RSS and trackbacks that blogs provide. The learning materials thus are only referenced and build up a network over the PLEs of all connected learners.

A further concept commonly found in PLEs is *tagging*, i.e. assigning freely selectable keywords (*tags*) to resources [140]. Tags serve as a means to create a structure that is non–hierarchical, allowing quick finding and retrieval of resources and enabling navigation paradigms like *tag clouds*. The use of tagging is widespread on the web, digital objects like images, videos, e–mails and web resources can be tagged in different applications. In PLEs, tags allow learners to categorize resources according to their own needs without a strict categorization schema like LOM [91]. Thus, in PLEs, the light–weight tagging replaces the metadata categorization of LOs (cf. section 2.2).

2.4 Self–directed Resource–Based Learning and Learning Resources

With the growing importance of the Web as a source for learning materials and the means to connect learners, learning paradigms that do not take into account learning in institutional educational settings exclusively have increasingly become prevalent. These learning paradigms shift the focus on learning where the learner herself is seen as an autonomous being, organizing and planning her learning process independently of a learning authority, like a teacher or tutor.

2.4.1 Self–Directed Learning

For example, the concept of Self–Directed Learning (SDL) [104] highlights the responsibility the learner takes when organizing her own learning processes without the authoritative support of a teacher or tutor (although teachers can take effective roles in SDL, but this is not the common case). Hiemstra [89] describes the learner as the central component in SDL: as learning can (but does not necessarily need to) take place in isolation from teachers and other learners, the learner has to take more responsibility for decisions concerning her own learning process. Thus, SDL emphasizes the learner's ability to control the context of her own learning processes including learning methods, activities and resources. Baumgartner [17] further touches the subject of the balance between autonomy and social action, stating that learning communities do have a place in SDL despite its focus on the autonomous individual.

2.4.2 Resource–Based Learning

According to Rakes [155], RBL is a style of learning that involves self–directed learning by using resources rather than by class exposition, and can be seen as a specific form of SDL. RBL is described as a learning mode in which the student learns from her own interaction with a wide range of learning resources. The teacher, not being necessarily present at the learning scenario, has to fulfill a role that involves giving support (and, in institutional settings, possibly choosing the resources to learn from), but is not interfering with the actual learning process. Thus, the teacher's role can be seen as a tutor who gives advice or feedback only on demand. With the omission of the role of the teacher, Breivik and Rakes [37, 155] identify *information literacy* as a crucial competence a learner has to have. Information literacy applied to RBL involves multiple process steps:

[8] `http://www.elgg.org/`, *retrieved 2011-03-16*

1. . . . knowing *when* there is an information need. For example, when a task, problem or issue occurs, the learner has to evaluate whether she already knows all needed information.
2. . . . *identifying* and *locating* the needed information.
3. . . . *evaluating* the relevance of the found information.
4. . . . *organizing* the information and make it accessible for future use. This encompasses recording relevant information and documenting their source. Further, relevant information should be structured "according to some logical pattern" [155].
5. . . . *using* the information efficiently to address the identified information need.

Whereas in RBL, the role of a teacher can still be existent, there are learning styles that assume a maximal autonomy of the learner, thus omitting the role of an authority like a teacher or tutor, or where this role is supplanted by a community. For example, Lindstädt et al. [120] define Work–integrated Learning (WIL) based on the observation that in many working scenarios a seamless integration between working and learning exists (referring to the notion of a *knowledge worker* [64]). Thus, they understand learning as the acquisition of knowledge and skills as a function or outcome of participation in authentic tasks. As working is social in nature, this includes direct or indirect support and guidance by other persons more experienced or more skilled [115]. Thus, the role of a knowledge worker is embodied in her interactions, sometimes acting as a learner and sometimes acting as a teacher or expert — depending on her experience and the task that is currently executed [191].

However, in this thesis, it is assumed that the existence of an authority is still possible, e.g. a tutor or expert may be integrated in the learning process of a group of learners and may occasionally give hints and recommend learning materials. Alternatively, learners can embark on their learning process completely autonomously. In order not to confine the supported learning types too much, the notion of RBL is chosen as a base theory that enables other learning paradigms. However, the need for a certain amount of self–direction is assumed.

2.4.3 Learning Resources

In a RBL setting, the notion of LOs is not adequate, as it is conceptually too imprecise. Further, it is a term that is very controversial, as different authors denote differing aspects of LOs that they consider relevant. For example, a LO implies the use of a certain structure, a clear learning intention and the systemic and institutional background. Thus, in this thesis, the term *Learning Resource (LR)* is preferred for denoting the codified learning materials. According to Rensing et al. [159],

Definition *A Learning Resource is a digital resource used for E–Learning.*

This definition is very generic and similar to the interpretation of the term LO according to the IEEE [91], yet it has the convenience on being not as controversial as the notion of the Learning Object. Further, it provides a multi–granular view on LRs. It covers fine–granular resources like media objects or snippets of single web resources as well as conventional LOs or whole courses of WBTs consisting of multiple documents. Further, a LR does not necessarily attribute the containment of information to a resource. For example, the start page of a social bookmarking application represents a starting point to a rich source of information, and therefore is a valid LR although it does not contain learning materials *directly*.

Further, the notion of LOs is usually understood to be afflicted with context, as it should be didactically and structurally self–contained. A LR does not require to be firmly rooted in an instructional context, and, in self–directed learning settings, the learner herself designates the embedment of a LR in her personal knowledge hierarchy and usage context.

Thus, through–out this thesis, this terminology is used for denominating content that can be used for TEL.

2.5 ELWMS.KOM — A System Supporting Resource–Based Learning with Web Resources

In this section, a short overview of a system named E–Learning KnoWledge Management System (ELWMS.KOM) [176, 125] is given that has been designed to support the scenario that is targeted in this thesis, which is RBL with web resources (cf. section 1.2). First, a model is introduced that identifies the important process steps of RBL. Then, the design goals for ELWMS.KOM are given and the implementation of ELWMS.KOM is briefly presented.

2.5.1 A Model for supporting Processes in Resource–Based Learning

Based on Tergan's model of process categories of individual knowledge management [187], Böhnstedt [28] presents a novel model of the processes in RBL (cf. figure 2.4) that incorporates five different process building blocks and is aimed at supporting the analysis and conceptualization of an implementation of RBL in the context of web resources. It is based on observations and feedback that were obtained in a user study with ELWMS.KOM.

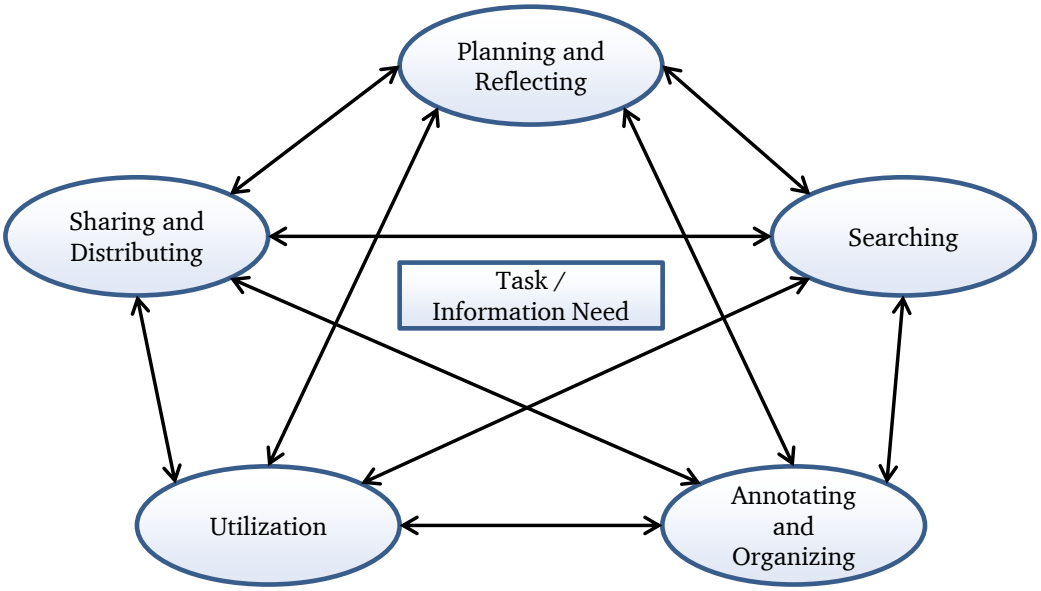


Figure 2.4: A model of the processes in Resource–Based Learning (cf. [28]). Five process building blocks in the learning process are depicted in the context of a learning task or an identified information need.

The five process building blocks are executed in the context of a *task* or identified *information need*, which can be interpreted as a learning goal.

Planning and Reflecting This process involves the analysis of a given task and transforming this task into a learning goal. Then, the learner defines a suitable course of action in order to achieve the goal. This includes an analysis of already known information and an identification of knowledge gaps as well as temporal constraints. The overall goal is partitioned into smaller sub–goals that represent realistic milestones on the way of attaining the learning goal. During the execution of the task,

the advancement is observed by the learner herself and adjusted with the set goals. After having achieved a (sub–)goal, the learner reflects on her learning process so far and readjusts the overall goal and adapts her approach to achieving the goal if necessary. Basically, this process step employs metacognitive processes that are defined in the theory of Self–Regulated Learning (cf. chapter 6).

Searching The process of searching is triggered when an information need has been identified. The learner decides which search strategy is likely to yield results effectively and efficiently. This process encompasses searching in resources that have already been found in prior RBL sessions and searching on the Web for relevant resources, often using a search engine or already–familiar information hubs (i.e. digital libraries, Wikipedia or social bookmarking sites like Delicious[9]). After a resource has been inspected, the learner rates its quality and its relevance for the learning goal. If the search process was not successful, the search strategy is adapted, e.g. by using different query terms or using other information sources.

Annotating and Organizing This process step encompasses all actions that involve saving, enriching and structuring found information. These actions include storing a resource on a storage medium, organizing found resources in a hierarchical folder structure, describing a resource with additional metadata or tagging in a social bookmarking application. Böhnstedt [30] shows that a majority of users store found resources on their hard disk, aggregate relevant information in a text editor or use the bookmarking functionality of their respective browser. Further, many users make paper notes or use e–mails for storing relevant information and some even employ most of these strategies. Writing an own summary about found information in a note or annotation already involves knowledge acquisition, as the information is immediately transformed to the understanding of the learner. This process results in an own "knowledge base" that is specifically tailored to the needs of the respective learner. Therefore, this process should involve creating an appropriate structure so that the found information can be efficiently accessed again later.

Utilization The utilization of resources includes all actions that are executed using the information contained in them, primarily geared towards completing the learning task. For example, if the learning task is learning a new programming language, a found code example can be analysed in order to understand a certain concept of that language. Or, even just reading found resources and trying to understand, or applying information in a research paper or a presentation can be understood as a utilization of resources.

Sharing and Distributing As learning is often residing in a social context, this process step is important for enabling the exchange of information. It involves any strategies that allow learners to communicate found information (or information about where relevant information can be found) to other learners that possibly have a similar learning goal. This process can include one–to–one communication (e.g. sending the URL of a resource via e–mail to a colleague), one–to–many communication (e.g. writing a blog post about best practices or solutions to a problem the learner had) and even many–to–many communication (e.g. collaboratively editing a research paper in a wiki). Böhnstedt explains that in her study, a large majority of the participants state that they would like to benefit from the research results of colleagues, though most find the exchange of resources too cumbersome or do not know who possesses relevant information.

As shown in figure 2.4, each process of RBL interacts with all other processes and a strict sequence of process steps does not exist. For example, results of a search process can immediately be shared or distributed without saving or organizing the found information. Further, if enough knowledge about

[9] `http://delicious.com`, *retrieved 2011-01-18*

the task is already present, the searching and organizing steps can be skipped. However, each of the respective steps is relevant in RBL and commonly encountered.

2.5.2 Design Decisions of ELWMS.KOM

This model of the processes in RBL describes a set of desiderata that an application to support RBL should optimally fulfill. In the following, design decisions derived from this model, an analysis of the target group and related work presented in the preceding sections are stated that ELWMS.KOM should conform to.

1. ELWMS.KOM is specifically targeted at self–directed academic learning settings and the role of the so–called *knowledge worker* [64]. Knowledge workers perform predominantly intellectual tasks and deal with information and knowledge. Four different types of knowledge work can be distinguished [108]: collecting information, analysing information, processing information and communicating information. ELWMS.KOM has been designed to support especially the collection and communication of information. A certain level of technology skills and frequent use of the Web as a source of information are assumed.
2. As the WWW becomes a major source for LRs, ELWMS.KOM strives to primarily support information search in the Web and organization of web resources. However, often only a fragment of a web resource is relevant for a learner and the rest of the web resource is boilerplate content that is not needed at all. ELWMS.KOM should allow the learners to store only *parts of web resources*, enabling them to create their personal knowledge base containing only relevant information.
3. Non–obtrusive accessibility should be provided. For a working flow without unnecessary interruptions, it is crucial that the learner can integrate this system into her common learning and research patterns. As ELWMS.KOM focuses on web resources as LRs, a tight integration of ELWMS.KOM into the "window to the Web", the web browser, makes the experience of the process more seamless.
4. A learner should not be enforced to organize relevant LRs in a way that does not reflect her own mental image of the found information. Therefore, in concordance with Hug and Friesen [94], ELWMS.KOM supports an *emergence model*, organizing LRs in a flexible way without restricting the learner to stick to a certain structure. This can be achieved by *tagging*.
5. Metadata support is important. As described in section 2.2, metadata are important for enabling an efficient retrieval of LRs. The different metadata categories defined by LOM [91] can be partially mapped to an informal learning scenario, but all metadata describing the authoring process or the didactic framework are irrelevant for web resources. For example, as there usually is no intended didactic function of the LRs, the LOM category *Educational* is not applicable. Further, LOM consists of over 70 metadata entities, and often users are overextended with providing such an amount of data, as specifying metadata often is complex and requires expert knowledge and competencies [38, 16]. Nevertheless, the learner should be enabled to attach as much and specific metadata if she sees the need to do so.
6. A problem that is often encountered with plain text tags is that these tags are often ambiguous. For example, plain text tags do not allow differentiating between the French city of *Paris* and the mythological Trojan *Paris*. Or, if a resource has been tagged with *Java*, is the topic of the resource about coffee, the Indonesian island or the programming language? In order to support disambiguation between tags, typed tags should be introduced that allow complementing a plain text tag with an additional semantic dimension by specifying *what kind of thing* this tag represents.

Böhnstedt [29] has shown that the tag types *Topic*, *Location*, *Event*, *Person* and *Type* are the most relevant for a RBL scenario.

7. The community aspect is of utmost importance, as learning processes are often social in nature — often necessarily due to the lack of a teacher or tutor. Thus, information exchange should be simplified and the LRs of a learner should be available for all other learners. This requires having a central storage of persisted web resources and the structure they are stored in. However, this community aspect is not the focus of this thesis, for further information see [29].
8. Providing support in the search process is vital. As there usually is no teacher or tutor in informal learning settings, the organization of the learning process is in the hands of the learner. This involves the execution of multiple metacognitive processes on the side of the learner: identifying the information need, setting learning and research goals, monitoring the level of completion of the current search and, eventually, reflecting on the learning process. In formal learning settings, the learner is guided by a teacher or tutor who structures the learning process accordingly. In informal settings, however, these processes have to be executed by the learner herself in a self–directed way. Thus, a system supporting RBL should support these processes adequately.

These design decisions have been taken into account in the implementation of ELWMS.KOM.

2.5.3 Implementation of ELWMS.KOM

ELWMS.KOM is a platform for supporting self–directed RBL. It focuses on web resources as LRs, and therefore is implemented as an add–on to the web browser Firefox[10], as commonly the web browser is the "window to the Web". It allows learners to store whole or partial web resources and to build a knowledge network by employing semantic tagging. A knowledge network is a structure based on semantic networks, which [183] describes as "a graphical notation for representing knowledge in patterns of interconnected nodes and arcs". Semantic tagging is an extended form of tagging that introduces an additional *type* to plain–text tags. This allows learners to mark a tag as a special semantic entity, e.g. disambiguating the tag "Paris" as a *Location* with the mythological Trojan *Person* "Paris". Further, it allows assigning non–topical tags to resources that provide personal retrieval hints, e.g. stating on which *Event* a certain paper or talk was given like "EC–TEL 2009" or who recommended a certain paper like the *Person* "Dinsdale". ELWMS.KOM provides default tag types, but a learner can easily define her own tag types [29]. The tagging process creates relations between tags and resources, and co–occurrences of identically typed tags allow inferring over non–explicit relations.

Knowledge networks exist in two granularities, for one there is a *personal knowledge network* that represents the tags and resources of one learner. Further, there is the *community knowledge network* that encompasses the knowledge networks of all learners combined. As learning is often a social process (cf. section 2.3), this community knowledge network enables the interaction, retrieval and collaboration across personal borders.

For providing a central storage of knowledge networks, a server–based approach has been chosen for the implementation of ELWMS.KOM. It uses the graph database and back–end *K–Infinity*[11] for storing all users' knowledge networks. *K–Infinity* provides the underlying graph database that is accessed via the *KEM–API* and a web application (the *Knowledge Portal*) that enables users to view and browse their knowledge networks. ELWMS.KOM consists of a WSDL web service that connects to the *KEM–API* and

[10] `http://www.mozilla.com/firefox`, *retrieved 2010-11-17*

[11] `http://www.i-views.de/web/index.php%3Foption=com_content&task=view&id=13&Itemid=45&lang=de_DE.html`, *retrieved 2011-03-17*

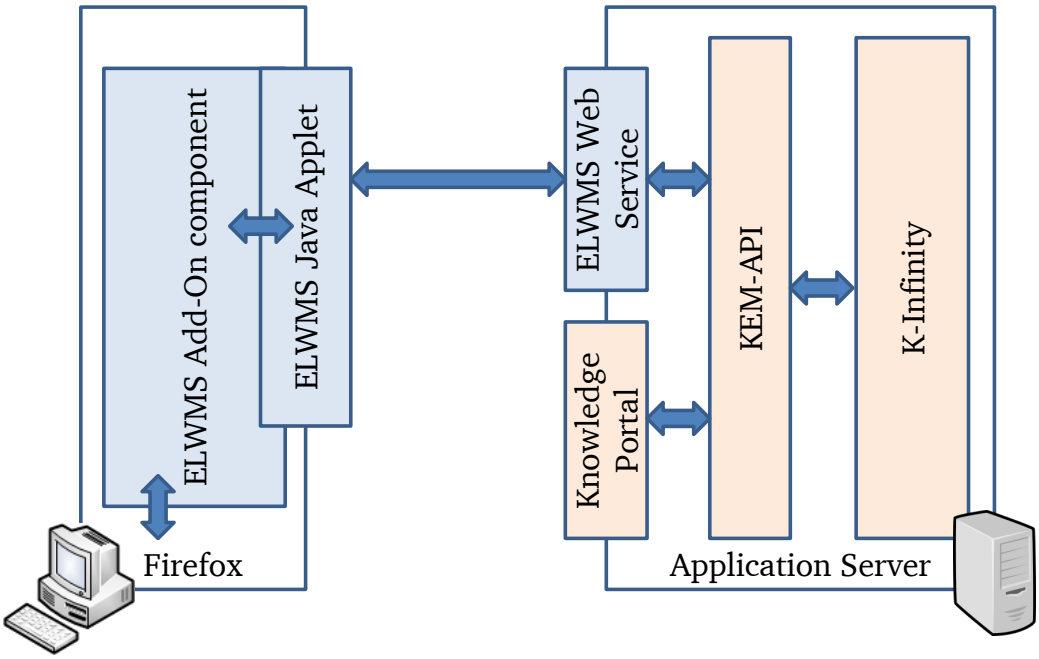


Figure 2.5: Architecture of all components of ELWMS.KOM. The server side components are provided by *K–Infinity*, the web service and the Firefox add–on are parts of ELWMS.KOM. The arrows denote inter–component communication.

the Firefox add–on programmed in XML User Interface Language (XUL) and Java that provides the user interface to the learner and communicates with the web service. Figure 2.5 shows the general architecture of all components.

ELWMS.KOM is integrated in the Firefox sidebar, therefore it provides an unobtrusive user interface that can be accessed while browsing without switching windows or starting a new program. The following relevant functionalities are accessible via the sidebar (cf. figure 2.6):

1. Importing a selected web resource (snippet). On importing, a window is opened where semantic tags can be assigned (see figure 2.7).
2. Recommendations. If the web resource that is currently browsed is already in the community knowledge network, in this area different recommended web resources are displayed. The recommendation process relies on structural properties of explicit relations between the current and recommended web resources.
3. The goal hierarchy. ELWMS.KOM allows learners to plan and structure their learning process by setting *Goals*. This is described in detail in chapter 6.
4. Activity stream. In this panel, small screenshots of the last persisted web resources are displayed. This enables learners to reflect on the resources they have already added and allows a quick access to resources that are relevant in the current learning episode.

Learners are enabled to set goals before starting a subsequent long–term learning process. These goals can be marked as “activated”, so that all persisted resources are automatically tagged with the active goal. This serves to quickly search for and collect relevant resources without having to manually assign them to the current goal. When a web resource is stored, the learner can add tags to the resource. Currently, there is a base ontology of tags supported by ELWMS.KOM that consists of the types *plain text*

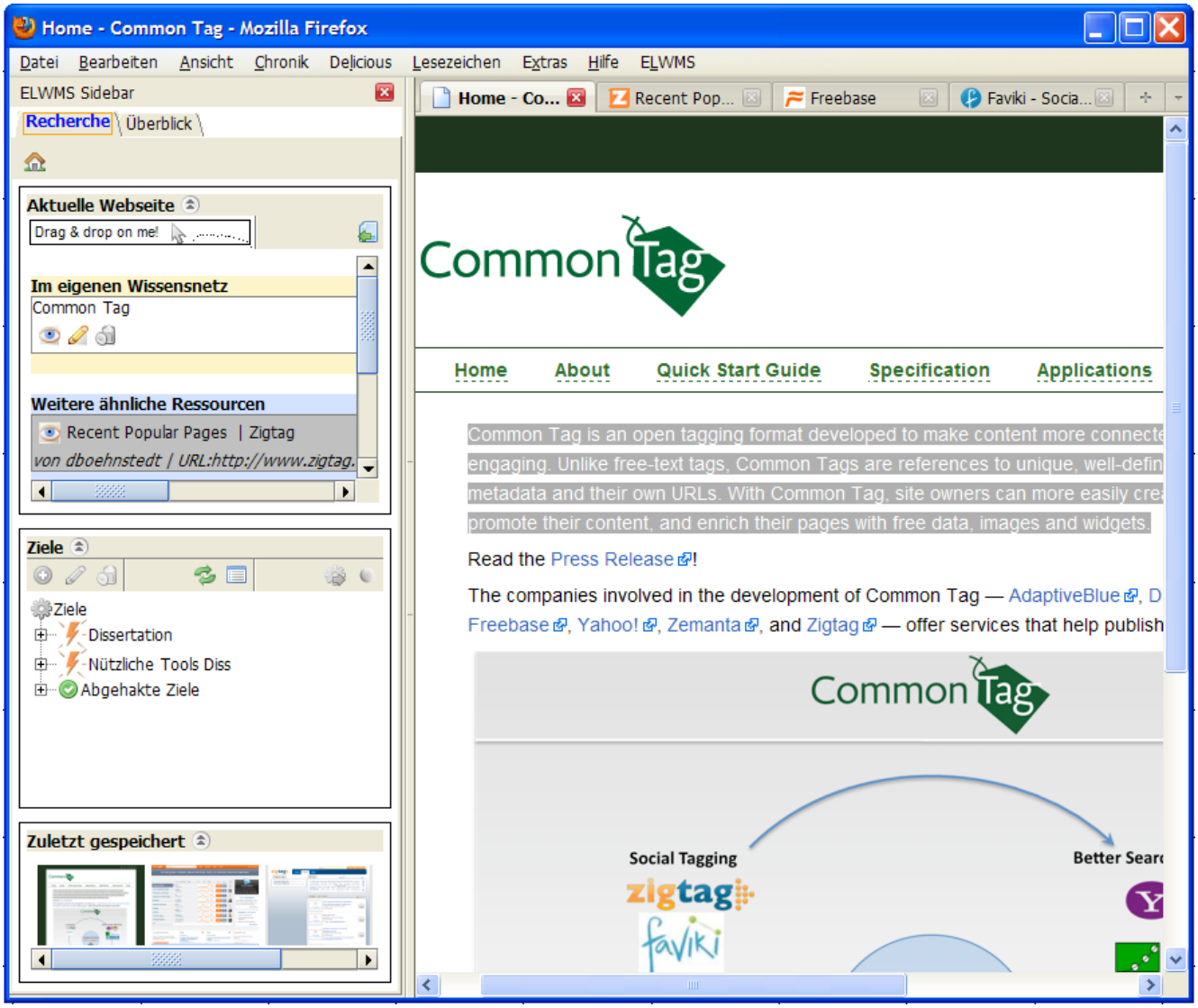


Figure 2.6: The ELWMS.KOM sidebar is embedded on the left side of Firefox. It displays (from top to bottom) import functionality by selection and drag & drop, a recommendation panel, the goal hierarchy and the activity stream.

Tags, *Topics*, *Persons / Organizations*, *Locations*, *Events*, *Type of the resource* (e.g. whether it is a *paper*, a *blog* or a *PDF* file) and *Goals*. For a discussion of these types, see [30].

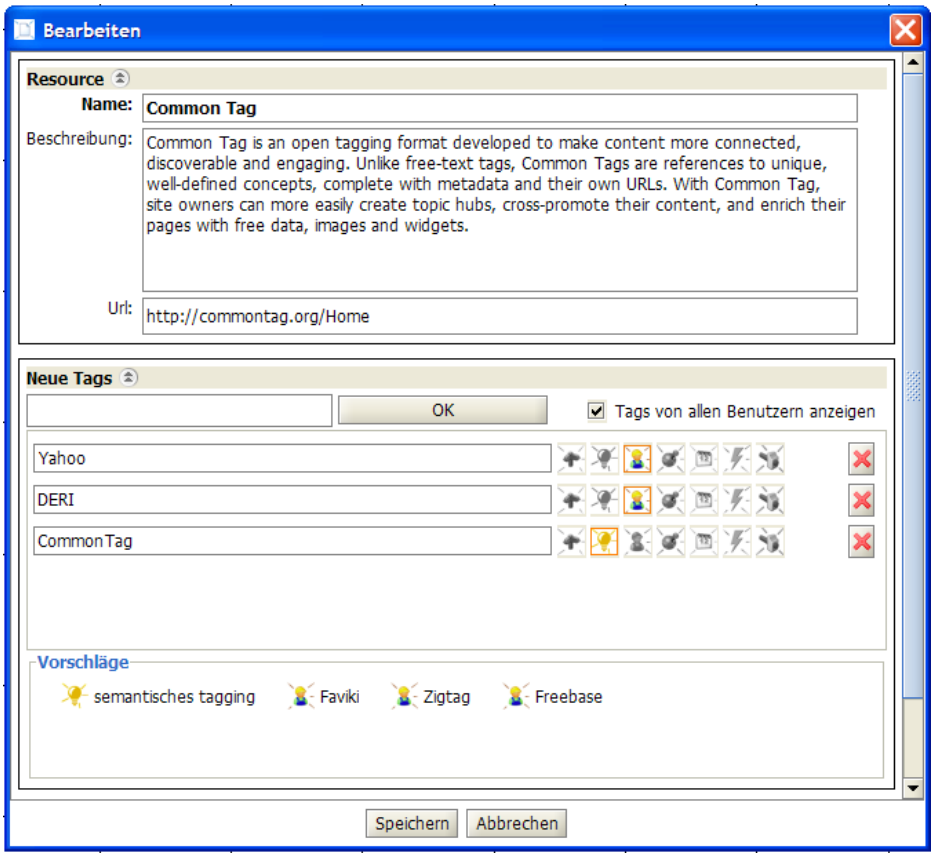


Figure 2.7: The web resource tagging view of ELWMS.KOM. The top panel contains the minimal metadata of a web resource: the title, a description or snippet and the URL. The lower panels contain the semantic tagging functionality and tag recommendations based on structural properties of the knowledge network.

For retrieval of the stored resources, learners can use the *Knowledge Navigator* to browse their personal as well as the community knowledge network using either the HTML view or a graphical display of the network structure. There are recommendations given based on the structure of the knowledge network and the ELWMS.KOM sidebar provides a listing of all web resources and tags per tag type. Export of the knowledge network is available in several formats; for example, scientific resources can be exported as BibTeX references and the goal hierarchy and all assigned snippets can be output as HTML.

ELWMS.KOM thus aims at supporting all process steps of RBL (cf. section 2.5.1) and constitutes the framework for all contributions of this thesis.

2.6 Conclusions and Discussion

The preceding sections have briefly outlined the differences between institutional learning and novel, self–directed learning scenarios and paradigms. Institutional learning settings that involve certification usually establish a specific framework for their learners, setting a curriculum, providing the technical and didactic infrastructure for learning like LMSs and providing learning materials in form of LOs. In non–institutional learning settings, however, as nobody takes up the role of the teacher or tutor, the learner has to organize her whole learning process autonomously, often without access to dedicated LOs.

Therefore, self–directed learners often utilize the vast amount of web resources as a source for learning materials. This has been presented as RBL using web resources and challenges in the support thereof have been shown.

For example, due to the lack of well–structured LMSs or content repositories, learners in RBL have to identify their information need, plan their proceeding in the learning process, search for relevant information on the web and store found web resources adequately in order to be able to retrieve them as needed. The system ELWMS.KOM has been presented in order to support these processes in RBL using web resources. It provides learners with a tool set for structuring their learning process, executing the information search and enabling learners to adequately organize and persist their learning materials derived from web resources by the application of semantic tagging.

With ELWMS.KOM, an exemplary learning process from the perspective of the learner consists of different steps:

1. Ideally, the learner identifies her information needs based on her learning tasks and plans her episode before starting a search. ELWMS.KOM supports this planned approach by allowing (and even prompting) learners to explicitly set goals and structure their advancement (cf. chapter 6).
2. The learner formulates a search strategy (e.g. by determining what kind of search is expected to yield results effectively and efficiently) and executes the search. Having found relevant web resources, the learner selects the part of a web resource that is important for meeting her current information need. Here, an automatic segmentation approach can improve the usability of this step and allow segment–wise retrieval (cf. chapter 4).
3. The learner organizes the found web resource segments in her personal knowledge network by assigning typed tags. Here, a consistent tagging structure with meaningful tag names and types is crucial in order to be able to retrieve the information again later. Therefore, the learner is supported by getting recommended automatically generated tags for special types (cf. chapter 5).
4. As learners may benefit of web resources that are already in the community knowledge network, ELWMS.KOM recommends other resources that are somehow related to the current information need. If the learner requests, the recommended resources may even be composed in other languages (cf. chapter 3).
5. Periodically during and after a learning episode, the learner is prompted to reflect on her learning process and respectively make modifications to the process.

The following chapters highlight the supported aspects of ELWMS.KOM.

3 Semantic Relatedness of Learning Resources

In Resource–Based Learning (RBL) settings, a major challenge for learners is finding relevant Learning Resources (LRs). As presented in chapter 2, there are several strategies to search for LRs that could be relevant for a certain information need. For one, the common way is using a web search engine or specialized digital libraries. In learning settings, however, where a community shares a similar context (like a learning group or a group of colleagues) does already exist, the probability that other members of the community already have found relevant LRs is high.

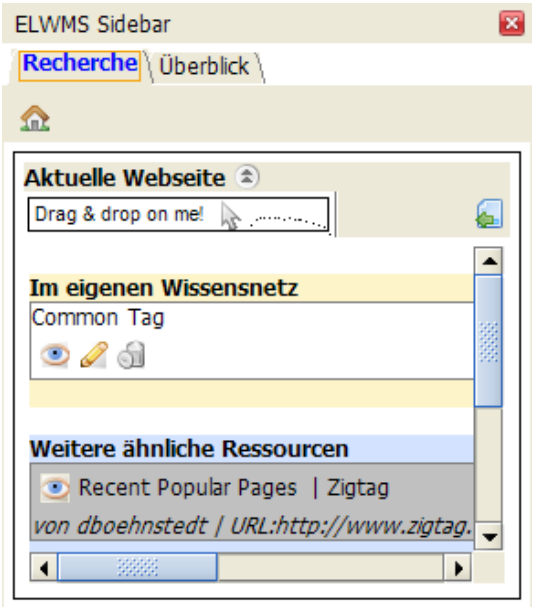


Figure 3.1: A recommendation is displayed in the left sidebar of ELWMS.KOM below the indicator that the learner has already saved the currently browsed web resource (about the open tagging format *Common Tag*). Another learner has saved a related web resource (about the tagging application *ZigTag*).

Therefore, the E–Learning KnoWledge Management System (ELWMS.KOM) uses a *recommendation engine* that attempts to provide information items like LRs or tags that are likely to be of interest to the learner (an example can be seen in figure 3.1). These recommendations bridge the gap between searching and sharing (cf. figure 3.2) on the basis of LRs already present in the knowledge network. However, up to now ELWMS.KOM only provides recommendations based on structural properties of the underlying knowledge networks, e.g. if there are explicit connections between two LRs over a defined set of tags. This means that if there is no explicit relation between two LRs, the recommendation engine is not able to infer this connection. Therefore, the formation of separate partitions of the knowledge networks is favoured, especially as different learners commonly use a different terminology for denoting related information (e.g. *TEL* and *e–learning*).

Another challenge that ELWMS.KOM has to meet with regard to recommending relevant items is that the overall knowledge network is expected to be sparse. In contrast to social bookmarking applications like Delicious, ELWMS.KOM does not have millions of users and therefore *collaborative filtering* [96] is not applicable for recommending items. Therefore, a *content–based* recommendation paradigm is targeted in this chapter.

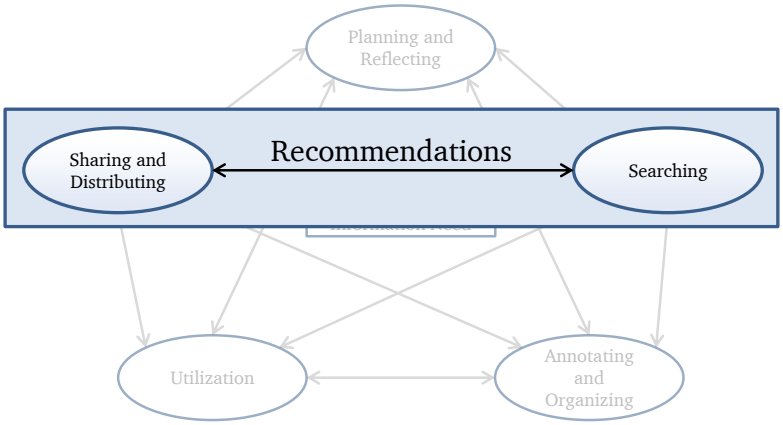


Figure 3.2: Supporting Resource–Based Learning by providing recommendations benefits the Searching and Sharing processes.

3.1 Introduction and Motivation

In this chapter, the challenges of providing *content–based recommendations* based on relatedness between tags and LRs are examined, which involve applying mechanisms of Information Retrieval (IR). It is a common task in Information Retrieval to find documents that are similar to a given query document. Content–based recommendation systems automate this step by providing similar result documents without the user's interaction or initiative. *Similarity* in this context has been usually determined as a measure of term overlap that occurs in these documents [8]. However, in recent work, a more high–level measure called *semantic relatedness* has been introduced that abstracts from the terminology used and aims towards a more semantic dimension, where the relatedness between *concepts* of the underlying documents is taken into account.

This is especially useful as humans tend to perceive similarity between documents based on *concepts* rather than on terms. In the context of Natural Language Processing (NLP), a concept is "an abstract or general idea inferred or derived from specific instances" [67]. Especially in domains where users need to find similar documents but do not exactly know the terminology, abstracting from terminology towards a more semantic measure is beneficial.

According to Budanitsky and Hirst [41], there is a considerable difference between the two notions of semantic closeness, semantic similarity and semantic relatedness. *Semantic similarity* denotes the degree of two different terms describing the same concept, e.g. the terms "cash" and "dough" have a high semantic similarity, because "dough" is a colloquial synonym for "cash". Further, the terms "building" and "bank" (in the sense of a bank building) have a considerable semantic similarity, as the concept "building" is a hypernym of the concept "bank" (i.e. it is a superordinate word encompassing the concept "bank"). However, there is no semantic similarity between the terms "cash" and "bank", as they describe completely different concepts. *Semantic relatedness*, however, denotes the degree of two different terms being related to each other, but do not necessarily describe the same concept, and therefore is more general than semantic similarity. For example, using the notion of *semantic relatedness*, the latter term pair is related, because the concepts "cash" and "bank" both occur in a common context. Thus, semantic relatedness mimics the associative perception of humans, taking not only the synonymy and hypernymy of terms into account but also other lexical relationships (e.g. meronymy and antonymy), functional relationships or frequent associations [139]. Therefore relatedness is the more general (broader) concept

since it includes intuitive associations as well as linguistically formalized relations between words (or concepts) [56].

Especially in the domains of TEL and RBL, different target groups with different levels of knowledge exist. For example, novices tend to be unaware of terminology of the domain they are learning, whereas experts are able to communicate in a brief manner using professional terminology. Further, in different stages of achieved expertise, different types of learning materials are important, giving either a broad overview or rather a very narrow scope of the learning domain.

Thus, for applications in RBL like ELWMS.KOM that support retrieval and recommendation of documents, being able to find semantically related documents is an essential task. The measure of relatedness is more suited to such a task than similarity, as learners do not only need to be recommended similar LRs about information they might already know but also related LRs that provide new insights or a novel perspective on the learning matter they work on. In the following subsections, the content of knowledge networks in ELWMS.KOM is analysed in greater detail and requirements for providing content–based recommendations in the specific scenario of RBL are highlighted.

3.1.1 Snippets

In a user study [174] with 64 participants, an evaluation was executed to see how learners select relevant content (for a detailed study description and a characterization of the participants see section 6.4.2). This user study served to examine how learners can be supported in organizing their learning processes with web resources by setting goals. During the study, participants were asked to collect learning materials from web resources, learn with the assembled information and take a performance test afterwards. The participants were instructed to collect the information from the web resources that they deemed to be relevant for their learning tasks, allowing them to select content in the *desired granularity*. In this study, 1,357 different snippets from 104 participants were collected.

For comparing the properties of snippets with "normal" bookmarked web pages (as these serve a similar goal), Delicious, a social bookmarking service that allows storing relevant URLs online, was crawled in order to obtain a comparison corpus[1]. 1,004 HTML pages thereof were downloaded and, after stripping HTML–specific content like markup, compared to the snippets gained from the study.

The results (see figure 3.3) and further manual analysis show that snippets differ from whole web resources in some aspects:

- Snippets mostly deal with a specific, well–defined domain, usually covering only one subject. Web pages, however, usually cover more information. This is not surprising, as snippets only account for a selection of the relevant information based on a specific information need.
- On average, snippets consist of 120 terms, whereas web pages consist of about 1,600 terms.
- 70% of snippets are smaller than 100 terms, 70% of web pages are smaller than 1,000 terms.

Based on observations in this analysis of snippets, the following requirements for an approach to generate content–based recommendations in ELWMS.KOM can be derived:

- In short snippets, there are only few significant terms. A larger terminological context is not available, thus the approach will have to abstract from the term level.
- The approach should be stable and provide good results, no matter how long the snippets are.

[1] The respective data was obtained by sampling random web resources from the most recent bookmarks feed (`http://feeds.delicious.com/v2/rss/`, *retrieved 2009-08-12*).

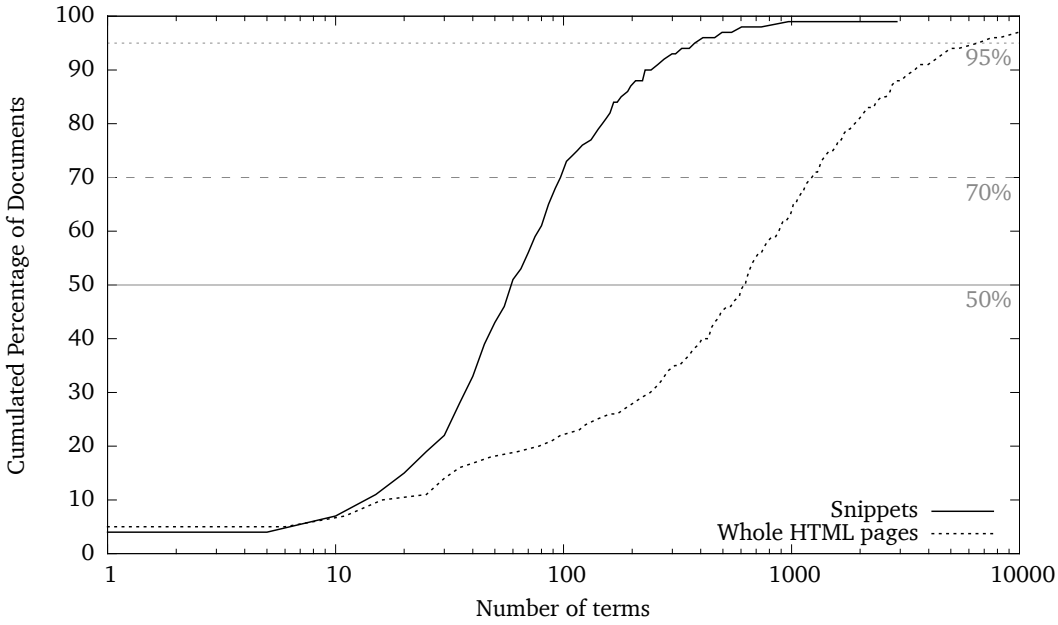


Figure 3.3: Cumulative term counts of snippets in comparison with term counts of full web pages. Snippets denote only the content the participants regarded as relevant to their current information need. This plot shows that snippets differ considerably from HTML pages in term counts.

- The snippets may be about any topic. Thus, the approach should be able to infer over arbitrary knowledge domains.
- Learners should be able to comprehend *why* two snippets are regarded as semantically related. This allows the learners to analyse if the recommended item is really relevant in their current learning situation.

Only German documents were collected in this user study [174]. However, in real–world scenarios, LRs often consist of documents in multiple languages.

3.1.2 Tag and Resource Language

In a second user study [28], 21 knowledge workers (3 female and 18 male participants with 7 being students of Information Technologies and 14 being members of research staff) used ELWMS.KOM in an academic setting over a period of several weeks. Before the participants started their research using ELWMS.KOM, they were given a short introduction into using the system. The participants were mainly using ELWMS.KOM in a research context, searching for relevant information about their respective field of expertise, which is a typical usage scenario in RBL. The participants were on average 30.6 years of age and used ELWMS.KOM in their daily work, storing relevant web resources they found in the knowledge network and tagging these resources accordingly. On inspection of the resulting web resources in the knowledge network, the following observations were made:

- In total, the participants stored 432 web resources. Although only 4 of the 21 participants are not German native speakers, a majority of 75.33% of the stored web resources are composed in English (see table 3.1). This is partly due to the academic setting, as publications are usually written in English, but it can be observed that also web resources stored for private use are often

composed in English. Six participants stored resources only in one language and merely 22.18% of the web resources are composed in German. Notably, the four non–German native speakers were not responsible for the better part of English resources but showed a similar resource language choice as the Germans.

- In order to describe the web resources in the knowledge network, the participants used in total 977 tags and on average 2.21 tags per web resource. In average, tags comprise of 1.73 terms with 542 tags (55.78%) consisting of only one term. 30.70% of all tags are English and 18.73% are German terms (cf. table 3.2). An additional 16.07% of tags are English but are technical terms that are conventionally used in German, too. In the use of tagging language, German participants expectedly more often used German tags, whereas non–German native speakers more often used English tags. The disparity in the usage of dates and named entities as tags is a result of individual tagging organization strategies.

Language	Web Resource Count	Web Resource Percentage	by Germans(4)	by Non–Germans (17)
English	333	75.33%	73.31%	79.45%
German	98	22.18%	23.99%	18.49%
French	2	0.46%	-	1.37%
Page forbidden (403)	1	0.22%	-	0.69%
Page unavailable (404)	8	1.81%	2.70%	-
Total	442	100.00%	100.00%	100.00%

Table 3.1: Web resources contained in the knowledge network grouped by language and fraction of resource language chosen by Germans and non–Germans.

Type	Tag Count	Tag Count in %	by Germans (15)	by Non–Germans (3)
English	300	30.70%	25.40%	42.91%
German	183	18.73%	22.17%	10.81%
English but conventionally used in German	157	16.07%	16.15%	15.88%
Ambiguous (German and English)	32	3.29%	3.67%	2.36%
Mixture of English and German	5	0.51%	0.59%	0.34%
Named entity (uni–lingual)	240	24.56%	28.63%	15.20%
Date or year	60	6.14%	3.39%	12.50%
Total	977	100.00%	100.00%	100.00%

Table 3.2: Tags used for web resources in different languages in ELWMS.KOM user sample. Note that German native speakers and non–Germans were involved and the number of participants does not match the numbers in the resource language experiment, as only 18 participants applied tags to resources.

This analysis shows that in this real–world setting, the usage of LRs often crosses language borders. This is especially the case in academic settings where the scope of the work environment is partly international. However, this does not necessarily apply to all learning settings that are targeted by ELWMS.KOM and thus the applicability is strongly dependent on the respective usage scenario.

In scenarios where resources and tags are composed in different languages, though, this language gap has to be accounted for. For content–based recommendations, this adds an additional dimension of complexity, as the language of documents has to be taken into consideration. Further, this does not only apply to LRs, but also to the used tags. The learners' choice of tags is often influenced by the language a

LR is composed in. This means that the same learner could use different tags for the same concept, e.g. one user tagged related LRs with the English “visual” and the German “Visualisierung”. This adds to the aforementioned challenges.

3.1.3 Structure of this Chapter

The remainder of this chapter is organized as follows: in section 3.2, an overview of related work is given and mapped with the above–mentioned requirements. In particular, a foundational approach for determining semantic relatedness, Explicit Semantic Analysis (ESA), is presented and identified as applicable. Section 3.3.2 describes implementation details and the used evaluation methodology. Based on ESA, some measurements concerning performance and coverage of this approach are analysed in section 3.4. Section 3.5 explores the applicability of ESA on cross–lingual relatedness calculation and presents a novel cross–language mapping strategy. In section 3.6, novel extensions to the basic ESA approach that additionally utilize the rich semantic information that the reference corpus Wikipedia provides are introduced and evaluated. Eventually, section 3.7 presents conclusions, open issues and next steps.

3.2 Related Work

Most approaches to compare documents in practical scenarios, e.g. search engines, apply the Vector Space Model (VSM) [8] in combination with the cosine similarity for calculating document similarity. Thus, approaches based on the VSM have in common to quantify the term overlap between documents. Documents are represented by high–dimensional feature vectors derived from the terms used in the document. The similarity between two documents is modeled by the angle between the representing vectors. However, as the vectors are entirely based on features that encode the term occurrences in the document, VSM is not applicable in cases of documents that are *semantically related* but have little term overlap. Specifically, in some scenarios it is beneficial if similarity is not expressed by means of terms but over the meaning of a document. Documents intended for differing audiences or written by different authors (e.g. beginners vs. experts) may be composed using different terminology, e.g. using technical terminology or synonyms and hypernyms. For example, the sentences “Willows often grow on river banks.” and “Trees of the genus *Salix* prosper on the borders of streams.” denote the same fact, although they only share the term “on”. Thus, although these documents describe the same semantic concepts, the term–based similarity will be rather low. This is called the *vocabulary mismatch problem* [181] or, alternatively, the *vocabulary gap* [204].

In scenarios that are possibly subject to such a vocabulary gap, there is the need to abstract from terms used in a document towards a more semantic representation. Thus, *relatedness* of documents is not to be expressed via common terminology, but rather by usage of terminology in a common semantic and conceptual context.

3.2.1 Semantic Relatedness and External Sources of Knowledge

As semantic relatedness is a measure that operates in a certain semantic context, it is impossible to be calculated with only the documents to compare. Milne and Witten [134] state that “any attempt to compute semantic relatedness automatically must also consult external sources of knowledge”. Thus, all approaches to determine semantic relatedness utilize additional information by employing *reference*

corpora in order to provide additional general knowledge. In related work, many different corpora have been used. Most provide structured access to semantic properties of terms (e.g. WordNet [67], Roget's Thesaurus [128, 97]), whereas other corpora, like Wikipedia, represent the underlying semantics inherently in the documents they contain.

One of the most popular reference corpora is WordNet [67], a lexical network of English words. WordNet provides networks of synsets[2] that contain terms like nouns, verbs, adjectives and adverbs, each representing a lexical concept. The synsets are interlinked with a variety of relations (e.g. denoting homonymy or synonymy). Semantic relatedness based on taxonomic structures similar to WordNet has been applied by several researchers [162, 98, 148, 66]. For example, Resnik [162] introduces a measure of semantic relatedness that is based on his notion of *information content* which depends on the probability of occurrence of a term in relation to a given corpus. For example, a concept with a high information content is highly specific for a given corpus. Semantic relatedness builds on this, constituting of the information content of the concept that subsumes both terms in a taxonomy's hierarchy. However, Patwardhan et al. [148] argue that this measure is unreliable, as it takes only into account a lowest common subsumer's information content and does not include the original terms, thus generating the same relatedness value for all term pairs that are in the same taxonomy hierarchy. They state that the quality of semantic relatedness strongly benefits from additional semantic information like provided by WordNet. Jiang and Conrath [98] build on Resnik's approach by augmenting the information content of the lowest common subsumer with WordNet path length and corpus statistics. They merge a content–based, node–centric information content approach with a node–distance, edge–centric approach and apply those to the WordNet noun synsets. According to Budanitsky and Hirst [41], this approach performs better than other measures they compared.

Another popular reference corpus that has been used for calculating semantic relatedness is Roget's Thesaurus. Jarmasz and Szpakowicz [97] use it as a base to calculate *semantic distance* between terms based on the path length in the thesaurus graph. They convert the distance to semantic similarity by subtracting the path length from the maximally possible path length.

However, both WordNet and Roget's Thesaurus are well–structured and have to be manually maintained by experts. Roget's Thesaurus, for example, dates from 1805 (with an edition from 1911 in the public domain). WordNet contains 155,287 unique noun, verb, adjective, and adverb strings organized in 117,659 synsets[3] with little growth over the years. Although general terminology is contained in both corpora, they cannot keep up with the rapid evolution of knowledge nowadays as they have to be maintained manually by linguistic experts, which is both expensive and laborious.

3.2.2 Semantic Relatedness via Document Corpora

Another approach that has gained momentum for the calculation of semantic relatedness in the last years is Latent Semantic Analysis (LSA) [59]. LSA is an approach that uses a custom corpus of documents to abstract from the used terminology and derives inherent semantic concepts from textual data. So, with LSA, different terms that are used as synonyms or are commonly co–occurring are mapped into a single concept. Further, by mapping terms, the overall corpus dimensions may be significantly reduced, thus transforming the search space. This projection and reduction is achieved by applying a singular value decomposition on a corpus matrix and then truncating the least significant values. The most significant values reflect an approximation of the strongest latent concepts that represent the documents of the

[2] A *synset*, or *synonym set*, represents a collection of synonyms that can be used interchangeably.

[3] `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`, *retrieved 2011-02-01*, Version WordNet 3.0

corpus. For each given document, the content of this document can be described using these latent concepts. LSA, although being a stable approach that performs well, has some limitations regarding the requirements stated in section 3.1. First, the number of dimensions that are reduced has to be determined in advance. However, the grade of reduction heavily depends on the topics of the documents that are present in the scenario. Second, the resulting concepts are sets of terms that define a semantic concept, but often these sets are not easily interpretable by humans. In settings that require humans to judge the quality of relations between concepts, this can be problematic. Thus, in the presented scenario, LSA is not a viable option.

In recent research, the collaboratively created and authored, open encyclopedia Wikipedia has been increasingly used for IR related tasks (e.g. [186, 83, 202, 135, 131, 99] and many others). This is due to the sheer amount of available articles (over 1.1 million content articles in the German version and more than 3.5 million in the English Wikipedia as of December 2010, cf. table 3.3), with each article ideally representing a distinct *concept*. Additionally, Wikipedia provides further semantic information about the concepts described in articles, most notably links to related articles (called *article links* or *intrawiki links*), a (mostly hierarchical) category structure (*category links*) and links to corresponding articles in other languages (*Cross–Language (CL)* or *interlanguage links*). Another criterion that makes Wikipedia a suitable reference corpus for IR tasks is that it is constantly updated by the community to the current state of knowledge, e.g. new articles are added and old ones are adjusted accordingly.

	English Wikipedia	German Wikipedia
Number of articles (rank)	3,571,974 (1)	1,196,433 (2)
Number of article links	113.8 million	34.1 million
Number of categories	∼ 660,000	∼ 100,000
CL links to other languages	7.3 million	4.7 million
Growth in 2010 (in percent)	∼ 300,000 (8.4%)	∼ 110,000 (9.2%)

Table 3.3: Selected descriptive statistics about the size of the English and the German Wikipedias as of December 2010.

Wikipedia therefore provides a good reference corpus for NLP tasks and is frequently used in scientific research. For example, *WikiRelate!* [186] is an approach that computes semantic relatedness between terms. Given two terms to analyse, WikiRelate! searches the Wikipedia article names (called *lemmata*) for the terms and calculates the distances between found articles based on the articles' contents and the category structure of Wikipedia. As it only supports computation of semantic relatedness between terms, this approach is not applicable to documents [78].

Kaiser et al. [99] introduce *conceptual contexts* of documents as linkage graphs that represent the document and its relations. Basically, they map documents to Wikipedia articles and apply a weighting function that determines the article's relatedness to neighbouring articles based on in– and outgoing article links. After removing all concepts that are only loosely related, they calculate the relatedness measure of the documents by computing the similarity of the link graphs. Kaiser et al. show that their approach outperforms a state–of–the art syntactic search engine and state that Wikipedia's article graph is a valuable source of semantic associative information.

3.2.3 Explicit Semantic Analysis

A promising approach to calculating semantic relatedness called Explicit Semantic Analysis (ESA) has been proposed by Gabrilovich and Markovitch [79]. Here, documents are not represented by means of terms but by their similarity to concepts derived from a reference collection of documents. ESA is based on the assumption that in the reference document collection, an article corresponds to a semantically distinct concept. Thus, by comparing documents based on their terminology to all articles in the document collection that have been pre–processed by tokenization, stemming, stop word removal and a term weight metric, a vector is obtained that contains a similarity value to each of the articles. This process step is called *semantic analysis*. The vector, called *semantic interpretation vector*, abstracts from the actual term occurrences and thus represents a semantic dimension of that document. A major advantage of ESA is that semantic relatedness can be calculated for terms and documents alike, providing good and stable results for both modes [79].

Formally, the document collection is represented as a matrix M (called *semantic interpreter*) with the dimensions $n \times m$, where n is the number of articles and m the number of occurring terms in the corpus. M contains *tf–idf* document vectors of the articles. *Tf–idf* [8] is a commonly used measure of relevance of a term in relation to a corpus D, where the *term frequency tf* of term t_i for each document $d_j \in D$ and the *inverse document frequency idf* of all occurrences of term t_i are taken into account:

$$\text{tf–idf}_{i,j} = \text{tf}_{i,j} * \log \frac{|D|}{|d_j : t_i \in D|} \tag{3.1}$$

For calculating the similarity between the document and the corpus, the *cosine similarity measure* (3.2) [8] is employed. Analogously, two documents represented as semantic interpretation vectors can be easily compared by using cosine similarity again.

$$sim(d_i, d_j) = \cos(\phi) = \frac{d_i \cdot d_j}{|d_i| * |d_j|} \tag{3.2}$$

ESA is applicable to different reference corpora. Gabrilovich and Markovitch have used the Open Directory Project (ODP) as well as Wikipedia, showing that ESA using Wikipedia outperforms the ODP reference corpus. They state that Wikipedia is especially practical for ESA as each of Wikipedia's articles ideally describes one concept.

Further, Gabrilovich and Markovitch show that ESA using Wikipedia as a reference corpus outperforms other approaches like WikiRelate!, WordNet, Roget's Thesaurus and LSA [79]. Kaiser et al. [99] see ESA as a competitor to their approach using conceptual contexts, but they do not compare their approach to ESA.

For longer documents, Gabrilovich and Markovitch propose feature generation using a multi–resolution approach. This approach generates different interpretation vectors on different levels of detail of documents, i.e. on word, sentence and paragraph level as well as for the whole text of the document. The most prominent features of these levels are summed up and build the interpretation vector for the whole document. Gabrilovich [77] presents test results showing that ESA using Wikipedia articles performs better in a multi–resolution approach, but states that computing different semantic interpretation vectors on different granularity levels considerably increases the computational complexity.

3.2.4 Cross–Language Semantic Relatedness

Cross–Language IR is a vibrant field of research that has been targeted by many researchers. Recently, the notions of semantic similarity and semantic relatedness have been researched with regard to CL settings [177, 181, 54], but already prior to this, the challenge to find documents which are not exactly corresponding to the terminology used in a search query has been examined.

In CL IR, a query is provided as a natural language document or search terms and a system answers the query with matching result documents [145]. For this task, it is a huge benefit for an approach if it is able to cross language borders, especially with the Internet consisting of numerous documents composed in different languages. The value of CL IR is apparent if a query is specific to a certain region or culture. For example, for the query "What is the highest building in Darmstadt", the probability to find a document answering that question is higher in the language space of the language that is spoken in the city of Darmstadt, which is German[4]. But also in queries that cover general knowledge, including documents in other languages can increase the probability to find relevant documents.

In CL IR, commonly a translation engine based on a dictionary is employed in order to map the query language to the target language. For example, Mitamura et al. [137] use a translator between English and Japanese/Chinese in order to translate the query term–by–term, considering the part–of–speech and other grammatical properties. The IR process is performed in the respective target language. Mitamura et al. identify the usage of *zero–anaphoras*[5] as a major source of error, as translator engines usually cannot infer their correct references. Further they state that another common problem is the ambiguity of terms. Bos and Nissim [33] translate all documents from the target language into the query language, as they argue that the impact of a bad translation of the query affects a system's precision more than a bad translation of some target documents. However, this is not practical in most scenarios, as it involves translating all documents that are to be searched into all languages considered as query languages.

Cheng et al. [51] state that a large fraction of search queries cannot be translated by using standard dictionaries and therefore they employ the Web as a source of terms. They utilize the bilingual translations contained in many Chinese and Japanese web pages in order to infer the correct translation of a search query, or, if a correct translation is not possible, at least a translation that is semantically related. Cheng et al. base their translation heuristics on term co–occurrences, showing that this performs especially well with named entities. They conclude that combining a dictionary–based approach and their translation heuristics yields the best results.

Ferrández et al. [68] use the inter–language index of EuroWordNet[6] for the translation task, which is a language–independent concept index that aligns the synsets of WordNets in different European languages. They observe that a majority of 87% of queries in datasets made available by the Cross Language Evaluation Forum[7] contains named entities, which have to be treated differently. Some named entities have to be translated (especially country names or organizations, e.g. the English name of the North Atlantic Treaty Organization "NATO" is translated to the French term "OTAN"), whereas others (e.g. person names like "Steve Jobs") usually have to be left intact. As EuroWordNet does not contain many named entities, Ferrández et al. employ the CL links of Wikipedia for the translation of named entities. Person entities are recognized and not considered for translation, whereas all other entities are attempted to be translated using the inter–language index and, if not possible, Wikipedia is again

[4] Actually, in Darmstadt the dialect of Hessian is prevalent, but let's just disregard that.

[5] In linguistics, an anaphora is a back–reference to a previously used term, e.g. "he", "it" and "their". A zero–anaphora is a reference to a not explicitly named term that is easily derivable for humans from the context.

[6] `http://www.illc.uva.nl/EuroWordNet/`, *retrieved 2011-02-22*

[7] `http://www.clef-campaign.org/`, *retrieved 2011-02-20*

consulted. They show that by using Wikipedia for named entity translation, the precision of an IR task between English and Spanish improves by 50%.

The value of Wikipedia's interlanguage links has also been acknowledged by many other approaches. Kinzler [102] presents a method to create multilingual thesauri from Wikipedia by using intrawiki links, link anchors, category links and CL links. The first three features serve to build a monolingual thesaurus for each language which connects terms and concepts by adequate relations. Afterwards, the resulting concepts are merged into a multilingual thesaurus by exploiting the CL links between the concepts. Each inter–language concept pair which is bidirectionally connected via CL links is merged into one single language–independent concept. This overcomes inconsistencies in the CL linking and is therefore robust to missing links.

Moreover, semantic relatedness across language borders has become a focus of research in recent years. Schönhofen et al. [177] investigate the usage of Wikipedia for CL IR, aiming to query and retrieve English documents by German and Hungarian[8] queries. For that purpose they first do a "word-by-word translation by dictionary", yielding in many cases a large set of word pairs for a single word in the source and the possible translations in the target language. In order to overcome this issue, they first aim to maximize the bigram similarity between the different translation combinations of adjacent words, consulting statistics obtained from the English Wikipedia as a reference corpus in the target language. Then, the links between pairs of articles containing the two translated terms in the article title are used to rank the translations. After having obtained the ranks for the translation pairs, Schönhofen et al. combine both measures to a final rank which results in an order describing the most probable terms. Although this approach benefits from the networked structure of Wikipedia which mirrors the semantic relatedness of concepts, it is still a term based approach which does not take the global term distribution, a measure of global term relevance, into account.

Potthast et al. [152], focusing on automatic cross–lingual plagiarism detection, consider a language–independent concept space to which a document collection is aligned for each supported language via a one–to–one mapping. This requires a reference corpus which contains articles describing the same set of concepts in different languages. Therefore, only a subset of articles can be considered for the semantic relatedness computation in the case of Wikipedia usage. For comparison they evaluate their approach by using `JRC-Acquis` [184] which contains mostly EU legislative documents as a reference corpus. Because of their assumption of a bijective article mapping function and their restrictive usage of disjunction of all articles, the direction of their mapping does not matter to the results. In a cross–lingual information retrieval scenario, they obtain the top ranking of the desired parallel document for 91% of all queries.

Sorg and Cimiano [181] present a slightly more elaborated approach which does not assume a one–to–one mapping between articles in the corpus but a many–to–one mapping for articles in the source language to articles in the target language. So each target article might be targeted from different articles in the source language. In their approach, they first compute the ESA interpretation vector in the source language and map it to the target language afterwards by summing up the relatedness values from all concepts in the source language pointing to a single concept in the target language. Their evaluation with different settings shows that ESA is able to match up to 46% of the results correctly. Further, they claim that for their scenario the best performing interpretation vector size is about 10,000 dimensions and state that a further increase of dimensions does not result in better accuracy but even decreases the accuracy.

Further, Sorg and Cimiano [182] aim to automatically detect missing CL links by using Chain Links information as classification features. Chain Links are defined for two articles $a_1^{l_1} \in W^{l_1}$ and $a_2^{l_2} \in W^{l_2}$

[8] cf. `http://en.wikipedia.org/wiki/Dirty_Hungarian_Phrasebook`, *retrieved 2011-02-27*

by relating the two Wikipedias in different languages W^{l_1}, W^{l_2} over an article pair that is linked by a CL link:

$$a_1^{l_1} \xrightarrow{\text{article link}} b_1^{l_1} \xrightarrow{\text{CL link}} b_2^{l_2} \xleftarrow{\text{article link}} a_2^{l_2} \tag{3.3}$$

where $b_1^{l_1} \in W^{l_1}, b_2^{l_2} \in W^{l_2}$. Their hypothesis is that "every article is linked to its corresponding article in another language through at least one chain link" and they show on a subset of 1,000 Wikipedia articles that this hypothesis matches for 95.7% of this subset. For classification, Sorg and Cimiano utilize graph based features and text–based features to select the best matching article in the other language from the candidates derived from article $a_2^{l_2}$'s Chain Links. They show on a selection of German Wikipedia articles without CL links that 81% of the proposed CL links are indeed valid and 92% of the results are at least related.

Another cross–lingual approach based on ESA is presented by Hassan and Mihalcea [86]. It differs from basic ESA in three points. First, it uses the Lesk metric instead of the cosine similarity. The authors argue that it places more emphasis on the overlap of the vectors than on the concrete values of the entries. Further, they do not make use of *tf–idf*, but replace *idf* by the logarithm of the inverse relative vocabulary size of the article correspondent to the represented concept. Finally, they make use of Wikipedia's category graph: the closer an article is to the root category, the more emphasis is put on its weight. Their evaluation with word pairs from the `WordSim353` dataset [70] translated to Spanish, Arabic and Romanian shows similar correlations for the monolingual English ESA and the original ESA. For cross–lingual ESA, their results on this dataset do not achieve the English monolingual correlation precision. For languages with less Wikipedia articles, the results improve by applying cross–lingual techniques in combination with the English language space. They explain this improvement with the better term representation by the usage of the large English Wikipedia concept space.

3.2.5 Summary of Related Work

All the presented approaches show that semantic relatedness is a promising field of research and that it does not stop at monolingual borders. Especially ESA has shown to yield good results on semantic relatedness in mono– and cross–lingual settings, thus its application is especially interesting for the scenario given in section 3.1.

Although ESA is commonly used with Wikipedia as reference corpus, it is not necessarily restricted to it. In theory, all textual corpora that follow the structure of providing unique documents (i.e. covering different topics) could be applied. Gabrilovich and Markovitch [79] apply ESA to a corpus derived from the ODP themselves, mapping concepts to the categories of the directory. Notably, Anderka and Stein [3] dismiss the hypothesis that the reference corpus needs to be semantically well–structured, i.e. semantic concepts are only described by one document. They show that ESA using the Reuters Newswire corpus and even random corpora may achieve comparable results to ESA using Wikipedia. Still, as Wikipedia provides distinct semantic concepts as labels (i.e. the lemmata of the articles), it is better for humans to interpret and understand the relatedness between documents.

Thus, in general, ESA fulfills the requirements as a *foundation* for the recommendation algorithm stated in section 3.1. ESA can cope with documents of arbitrary size, has the backing of a broad knowledge base (in this case Wikipedia) and performs well compared to other approaches. Therefore, in the following sections, the applicability of ESA is explored in more detail.

3.3 Implementation of ESA and Evaluation Methodology

After ESA has been introduced in section 3.2.3, section 3.3.1 briefly shows implementation details of ESA as it is applied in this thesis. Further, in section 3.3.2, different evaluation corpora and methodologies are presented that are comparable with the requirements given by the scenario given in section 3.1.

3.3.1 Implementation of ESA

In this section, ESA is briefly revisited and the process of creating a reference corpus from Wikipedia is presented. The overall process is shown in figure 3.4.

- First, a Wikipedia dump[9] (in this thesis, the dump of the German Wikipedia from June 2010 is used unless noted otherwise) is pre–processed with stemming, stop word removal, article filtering, *tf–idf* calculation and normalization. There are different parametrizations with regard to stop word removal and article filtering that are analysed in detail in section 3.4.
- Then, all article vectors derived from the first step are aggregated into the semantic interpreter matrix M (also only called *semantic interpreter*) with the shape $n \times m$, where n is the number of articles and m the number of terms. M is highly sparse, for a semantic interpreter including all articles and terms it is typically filled at about $0.1 - 0.6‰$. Therefore, it is internally stored in an efficient sparse matrix format, where only the non–zero entries in the matrix define the actual size of M in bytes.
- For each document d that is to be compared, the same pre–processing steps have to be executed, so that the result is the document vector v_d with the form $m \times 1$, where m is the number of terms.
- As all document vectors are normalized, the interpretation vector $i_{esa} = M \cdot v_d$ that represents the cosine similarity of v_d with all article vectors of M is simply computed by applying the inner product with the matrix M.
- Finally, the result is the interpretation vector i_{esa} with the dimensions $1 \times n$.

This interpretation vector i_{esa} is the representation of the original document in terms of Wikipedia concepts included in the semantic interpreter. Therefore, it is the foundation for all further approaches that calculate the semantic relatedness. For example, the semantic relatedness of two documents can be determined by applying a vector–based similarity measure. Analogously to [79], in this thesis the *cosine similarity* is used.

3.3.2 Evaluation Methodologies and Corpora

There are different scenarios that semantic relatedness has been applied to in related work. Depending on the use case of the respective approaches, there are several evaluation designs and evaluation corpora that can be applied.

In the following subsections, different corpora are presented that have properties that are relevant for the given scenario:

Term–based Relatedness In ELWMS.KOM, tags are usually consisting of one to two terms (on average 1.73 terms) and a majority of tags are represented by nouns. Therefore, a corpus that is comparable to the tags in ELWMS.KOM should consist of single terms (cf. subsection 3.3.2) or very short

[9] available from `http://dumps.wikimedia.org/`, *retrieved 2011-01-11*

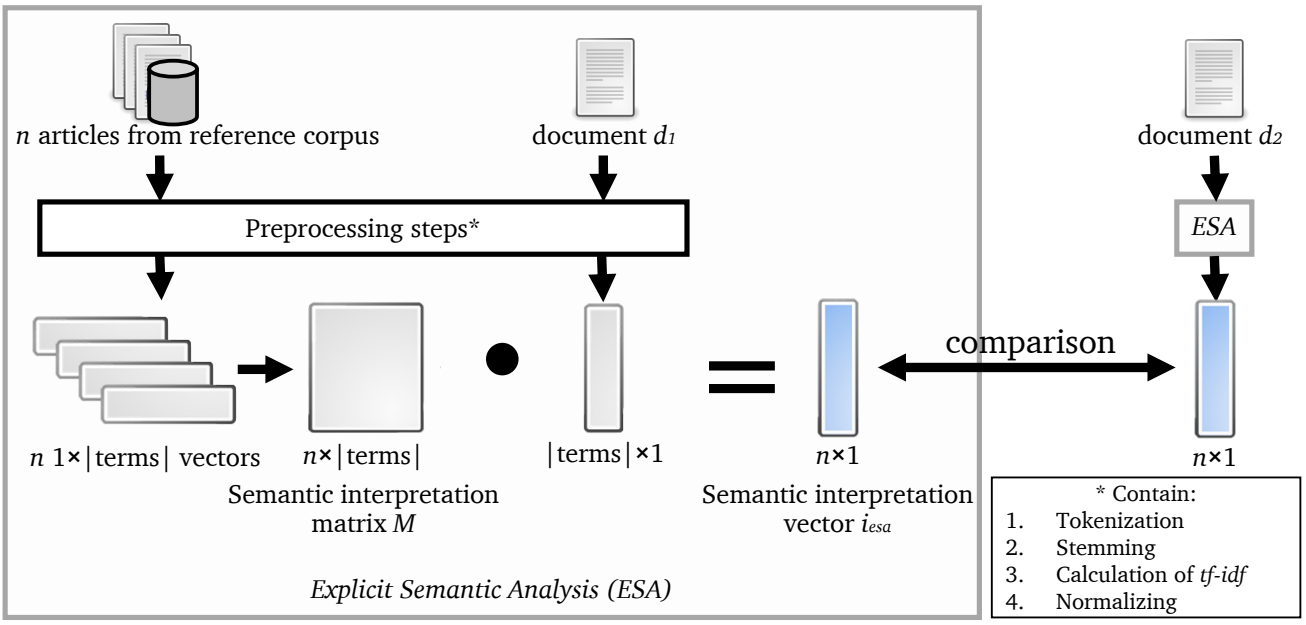


Figure 3.4: Process of creating a semantic interpreter from Wikipedia articles and deriving a semantic interpretation vector i_{esa}.

multi–term documents (cf. subsection 3.3.2). Another observation that was made in the tags of ELWMS.KOM is that some tags represent a generic knowledge domain (e.g. “e–learning”), whereas other tags are very specific and denote a clearly delimited concept (e.g. “mobile social search”). This should be reflected in the specificity of the corpora’s terminology. Therefore, several corpora with different properties have been used to examine the applicability of semantic relatedness.

Document–based Relatedness Snippets in ELWMS.KOM are usually short and contain on average 120 terms. A corpus that is used to measure the quality of a semantic relatedness approach thus should reflect this document size approximately. Further, the learning intention of users is an important aspect in the choice of documents. As, to the knowledge of the author, no appropriate corpus is existing that reflects these requirements, the novel semantic corpus `Gr282` is presented (cf. subsection 3.3.2) that conforms to this specification.

Multilingual Tags and Documents As shown in section 3.1.2, users of ELWMS.KOM in the examined academic environment often collected resources and created tags in different languages. Thus, appropriate corpora should be used that allow evaluating a cross–language semantic relatedness approach for term–based and document–based relatedness additionally to the monolingual approaches.

The following subsections give an overview of the selected corpora and present the applied evaluation methodologies.

Relatedness of Term–Term Pairings

In scenarios applying semantic relatedness to word sense disambiguation [70, 148, 84], the common methodology to evaluate an approach is by comparing human judgements of relatedness of a set of term pairs to the ratings the respective algorithm has calculated. The focus of this evaluation approach is not the *absolute value* of the respective ratings but rather the *order / ranking* in comparison to the

human ratings. Therefore, usually a rank correlation measure is applied, for example *Spearman's rank correlation coefficient* (also called *Spearman's* ρ). This measure can be applied to two k–sized lists x and y containing pairwise values. It determines how good the correlation of the values can be approximated by a monotonic function. First, for each variable x_i and y_i, the ranks $rank(x_i)$, $rank(y_i)$ of those values are determined. In case of equal relatedness values (called *ties*), the average of the respective ranks is assigned. Especially the lower relatedness boundary of 0.0 is probable to occur several times, e.g. when no term overlap exists. After cleaning of ties, the Spearman's rank correlation coefficient ρ is defined [141] as

$$\rho = 1 - \frac{6 \sum_i^k \text{diff}_i^2}{k(k^2 - 1)} \tag{3.4}$$

where diff_i is the difference between ranks of x_i and y_i and k the size of the samples.

The significance of the difference between two correlations can be determined by using t_{diff} [69]. It is used to check whether the correlation between the pairs of variables (x, y) and (z, y) is significantly different. It is defined as:

$$t_{\text{diff}} = (\rho_{xy} - \rho_{zy}) \sqrt{\frac{(k-3)(1+\rho_{xz})}{2(1-\rho_{xy}^2 - \rho_{xz}^2 - \rho_{zy}^2 + 2\rho_{xy}\rho_{xz}\rho_{zy})}} \tag{3.5}$$

The resulting values for t_{diff} are compared with the critical values of the t–distribution.

Datasets for Monolingual Evaluations

There are several datasets that have been used in monolingual semantic similarity evaluations, most notably the Rubenstein and Goodenough [164] similarity dataset (called `Rub65`) encompassing 65 pairs of nouns that have been rated by 51 humans for their similarity. There is a German translation of this dataset `Gur65` [84] with 24 human raters, which unfortunately does not exactly correspond to the English version (cf. table A.1 in appendix A.1). For monolingual settings, these inconsistencies are not relevant, but for evaluating multilingual approaches this matters.

For determining semantic relatedness, the German dataset `Gur350`[10] provides 350 term pairs with their respective relatedness values given by 8 subjects (cf. table A.2 in appendix A.1). In comparison to `Gur65`, this dataset contains not only nouns but also verbs, adjectives and adverbs. Further, named entities are included in this dataset, e.g. *Benedikt*, *VW* and *Opel*. This makes `Gur350` a more challenging dataset that is not fully applicable to semantic relatedness approaches that use formal ontologies, e.g. WordNet.

Dataset for Multilingual Evaluations

The multilingual dataset `Schm280` [172] is adapted from the English `WordSim353` dataset created by Finkelstein et al. [70] (cf. table A.3 in appendix A.1). It contains 280 English noun pairs with their German equivalent translated by up to 12 participants, each value pair with a relatedness value rated by at least 13 subjects.

Relatedness of Query Term—Document Pairings

A common scenario for semantic relatedness is the task to find a related document for a given query term, for example in IR settings [203]. Usually, in evaluations for such a scenario, there are no rated

[10] No publication known, available at `http://www.ukp.tu-darmstadt.de/data/semantic-relatedness/`, *retrieved 2011-02-25*

term pairs but rather a query that has to be mapped to a correct document. This is a task that is much more demanding to the respective approach, as it must be able to calculate a semantic relatedness measure between a single term and a possibly multi–term document. Some approaches presented in section 3.2 do not support taking into account multiple terms (e.g. [186]). A corpus that is commonly used to evaluate such a setting is a "word choice problem" corpus, i.e. having a term as query that is rare and attempting to select the correct description from a set of multiple possibilities. A data set that is often applied here is the `TOEFL` corpus [113] that consists of a set of 80 query terms and for each a selection of four possible synonyms. Usually, this corpus is used to evaluate semantic similarity approaches, but it has been applied to semantic relatedness as well [203]. A German equivalent is the Reader's Digest Word Puzzle corpus[11] (`RDWP984`). It was obtained from the 2001 to 2005 editions of the German Reader's Digest Magazine and contains 984 multiple choice questions consisting of a query term and four possible answers in form of a single term or a short definition (cf. table A.4 in appendix A.1). This dataset contains highly domain specific, rare terminology and therefore is a challenging corpus for determining whether a semantic relatedness approach is able to provide a good coverage of the terminology.

The quality of a semantic interpreter with a reduced article set is indicated by two different values: *Coverage* and *Accuracy*. Accuracy can be differentiated in *Local Accuracy* and *Global Accuracy*.

Coverage denotes the ratio of queries for which ESA is able to calculate a result and the total number of queries. As ESA maps documents to concepts according to the term overlap of the concepts' articles, it is crucial that the terms used in the documents are reflected in the semantic interpreter. A query is considered as *covered* by the semantic interpreter, if any relatedness can be calculated between the query term and the descriptions. Thus, Coverage is an indicator of enclosure of the terminology that is needed in order to generate a result.

Local Accuracy is the ratio of queries answered correctly by ESA and the number of covered queries. For all covered queries, the answer is scored as correct that is most related to the query. Thus, local accuracy represents the quality of the evaluation without taking into account queries that could not be answered due to terms missing in the semantic interpreter.

Global Accuracy is the ratio of queries correctly answered by ESA and the total number of queries. It does not take into account the accuracy loss resulting from non–covered queries and represents a quality measure reflecting a real–world setting.

Relatedness of Document–Document Pairings

Comparing a query document to a set of other documents and finding the most related match is a typical task in Information Retrieval. For evaluating such an approach, a methodology is employed that is used to evaluate search engine rankings [47]. Basically, a semantic relatedness value is calculated for each document $d_q \in D$ and all $d_i \in D \setminus d_q$. The result is a list that is ranked by decreasing relatedness. d_q and a compared document d_k at rank k are defined to be semantically related (i.e. $r_k = 1.0$) if they cover the same or similar concepts. If documents are semantically related, they are allocated to the same semantic group D_q (equation 3.6).

$$r_k = \begin{cases} 1 & \text{if } d_q \text{ and } d_k \in D_q \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

[11] Available at `http://www.ukp.tu-darmstadt.de/data/word-choice-problems/`, *retrieved 2011-02-08*

Further, *precision at rank* and *recall at rank* (equations 3.7 and 3.8) are used to calculate the *average precision* (equation 3.9) over one relatedness comparison for different recall values.

$$precision(k) = \frac{1}{k} \sum_{1 \leq i \leq k} r_i \tag{3.7}$$

$$recall(k) = \frac{1}{|D_q|} \sum_{1 \leq i \leq k} r_i \tag{3.8}$$

$$average\ precision = \frac{1}{|D_q|} \sum_{1 \leq k \leq |D|} r_k * precision(k) \tag{3.9}$$

Depending on the properties of the used evaluation corpus, there are two different ways of presenting the results. In case that one document corresponds to exactly one other document, a *top–k* [8] approach can be used. In *top–k*, a precision value is given for the *k* highest ranked results for each document, e.g. if a corresponding document is returned at rank 2, the *top–1* result would be 0.0, whereas the *top–5* would yield 1.0. As result, an average over all *top–k* results is given. As there is only one relevant document, recall is always either 0.0 (relevant document not in *top–k* result set) or 1.0 (relevant document is in *top–k* result set).

Alternatively, if one document corresponds to a set of other documents, *top-k* is not reasonably applicable. Here, all pair–wise comparisons are averaged and the average precision is plotted against interpolated recall, resulting in a so–called *precision–recall diagram* [47]. A precision–recall diagram represents a graph of the trade–off between precision and recall. For summarizing the quality of such a diagram numerically, two measures are usually given: Break Even Point (BEP) and Mean Average Precision (MAP). The BEP [200] represents the point where precision equals recall (and, as shown in the plots given in section 3.6.4), the interpolated precision–recall curve crosses $f(r) = r$, i.e. the angle bisector of the first quadrant). The MAP is the average of the precisions that have been computed for all documents.

Dataset for Monolingual Evaluation

For scenarios that enable the functionality of document recommendation, an evaluation corpus was needed that meets several requirements:

- The evaluation corpus should consist of German documents, as the focus of this research is based on the German Wikipedia.
- Documents in the evaluation corpus should conform to the snippet definition given in section 3.1.1, i.e. a majority of documents should contain between 20 and 200 terms.
- Documents in the evaluation corpus should honour the scenario of RBL with web resources. That is, they should contain a narrow scope of topics and be basically appropriate to meet specific information needs.
- Documents should contain different topics and have different scopes, i.e. should not only represent narrow factual knowledge but also contain opinions and overview information, making it a challenging task for semantic relatedness approaches.

Thus, a novel semantic corpus called Gr282 (cf. table A.5 in appendix A.1) has been built in a user study. Eight participants (2 female and 6 male, four students of Information Science, two students of Educational Science and two research assistants) were asked to research answers to a catalogue of ten questions (for a full listing see appendix A.2) using only fragments of web resources. For each question they were to find five snippets that (partially) contained the answer to this question using one of four different search engines (*Google*[12], *Yahoo!*[13], *Bing*[14] and *Ask*[15]) in order to ensure diversity of found web resources. Further, they were asked to restrict the snippets' length to 20 to 200 terms. This was not a fixed requirement though, if needed, the participants were allowed to collect larger web resource fragments.

In order to conform to the fourth requirement named above, the questions were formulated in a way that five different types of questions were asked with each type featuring two questions. Following question types were identified as relevant for the given scenario:

- *Opinions*, e.g. "Is the term *Dark Ages* justified?"
- *Facts*, e.g. "What is the FTAA?"
- *Related snippets* to a common topic, e.g. "Find examples for internet slang!"
- *Homonyms*, e.g. "What are Puma, Jaguar, Panther, Tiger and Leopard?"
- *Broad topics*, e.g. "Find information about the evolution of man!"

Gr282	
Size of corpus	282 documents
Average length of snippets	95.21 terms
Minimum length	5 terms
Maximum length	756 terms
Standard deviation	71.31 terms

Table 3.4: Short descriptive summary of novel corpus Gr282

After having collected the answers, duplicate answers and answers from the same sources were discarded. Finally, the evaluation corpus consisted of 282 snippets (a short summary is available in table 3.4) that were labelled with their question types and manually split into groups of different semantic concepts. Because, as expected, homonyms and broad topics showed to be consisting of snippets with different meanings, different semantic groups could be formed for some questions (cf. appendix A.2). For example, for the question that asks about the meaning of "Puma, Jaguar, Panther, Tiger and Leopard" there are different correct answers. First, they all belong to the biological feline genus *Panthera*. Second, they are all project names for Apple's Operating System OS X. Third, they are all common names of war tanks (however, none of the study participants answered with this option). Thus, this question spans three semantic groups.

In the evaluation, an IR task was executed with the expectation of getting all semantically related snippets (i.e. all snippets in the same semantic group) before all semantically unrelated snippets. As the semantic groups do have different sizes, a *top–k* evaluation is not applicable. Therefore, in this thesis, *precision–recall* diagrams and BEP and MAP are given as the results of the Gr282 evaluation.

[12] http://www.google.de/, *retrieved 2009-10-02*
[13] http://de.yahoo.com/, *retrieved 2009-10-02*
[14] http://www.bing.com/, *retrieved 2009-10-02*
[15] http://de.ask.com/, *retrieved 2009-10-02*

Dataset for Multilingual Evaluation

The Europarl corpus [105] is a multilingual collection of sentence–aligned protocols of proceedings from the European Parliament in eleven languages. The documents consist of full, grammatically correct sentences in natural language grouped in approximately 4,000 chapters[16], translated by professional translators. A challenge for cross–lingual relatedness approaches is the occurrence of many named entities (e.g. speakers) and the variability of the translations.

Europarl300		
	English	German
Size of corpus	300 parallel documents	
Average length of snippets	28.08 terms	26.75 terms
Minimum length	4 terms	4 terms
Maximum length	111 terms	111 terms
Standard deviation	17.24 terms	15.77 terms

Table 3.5: Short descriptive summary of corpus Europarl300

Due to computational constraints (for this cross–lingual evaluation, each document has to be compared to all documents in the parallel language, resulting in n^2 comparisons) only a subset of the Europarl corpus was taken, containing 300 parallel documents in German and English, in the following referred to as Europarl300 (cf. table 3.5 and table A.6 in appendix A.1). This subset consists of the first 300 documents the Europarl test data of the second Workshop on Statistical Machine Translation 2007[17]. Because one document has exactly one correspondent document in the other language, a *top–k* evaluation is applicable in this scenario.

3.3.3 Conclusions

This section has given a short overview of the implementation of ESA and the used evaluation methodologies and corpora. In particular, the novel semantic corpus Gr282 was presented that corresponds to the snippet definition given in section 3.1.1.

The next sections cover an analysis of different properties of ESA. No related work has performed a comprehensive analysis of the impact of filtering certain articles or terms before building a semantic interpreter yet. Therefore, section 3.4 analyses different aspects of ESA in terms of semantic interpreter size, terminology coverage and quality with regard to article and term filtering. Here, article filtering strategies described in related work as well as novel article filtering strategies are examined. Further, exploiting the availability of the applied reference corpus Wikipedia in multiple languages, section 3.5 presents an approach to map semantic interpreters across language borders. Finally, adjustments to ESA called eXtended Explicit Semantic Analysis (XESA) are presented in section 3.6 that take into account the rich implicit semantic structure of Wikipedia.

3.4 Optimization Strategies for ESA

A challenge of ESA is its high computational complexity. For each document that is to be transferred into its semantic interpretation vector i_{esa}, the whole semantic interpreter M has to be multiplied. Thus, the

[16] Depending on the language, this number varies.

[17] http://www.statmt.org/wmt07/shared-task.html, *retrieved 2011-03-12*

performance of the approach is directly dependent on the dimension size of M, namely the number of articles and the number of contained terms. An appropriate reduction of M is beneficial to the overall performance (and therefore the applicability of ESA). As no related work has performed a comprehensive analysis of the impact of filtering certain articles or terms before building a semantic interpreter yet, this section analyses different aspects of ESA in terms of semantic interpreter size, terminology coverage and quality with regard to article and term filtering.

There are two basic possibilities to reduce the dimensions of the semantic interpreter:

1. Filter the number of articles. The assumption is that not all articles equally contribute to the semantic comprehensiveness of the semantic interpreter. There are several strategies to filter articles that can be applied to the semantic interpreter.
   - Removing articles that are very short. These articles contain only few words and often serve as a placeholder (called *stub*) for a more elaborate article. Usually, one single sentence gives a short description about the respective concept, but no further details are given. These articles often exhibit too little specific content and too little context in terminology to be reliably used for calculating semantic relatedness.
   - Filtering articles that are very specific or very general.
   - Removing articles that do not adhere to Wikipedia's "one article — one concept" paradigm (e.g. articles containing lists of authors ordered by name).
   - Filtering articles that describe a certain class of concept (e.g. a person).
2. Filter the number of terms. According to Zipf [206] the frequency of a term in a natural language corpus is inversely proportional to its rank in the frequency table. Figure 3.5 illustrates the distribution of terms in the German Wikipedia. The figure plots the number of a term occurrence in respect to its rank. For the German Wikipedia, the resulting distribution is only approximately Zipfian as the stemming of terms skews the distribution slightly. Few terms appear very often (called *stop words*) while a large number of terms appear only in a few documents (rare words) and form the *long tail* of the distribution. Removing these terms is a strategy that is often applied in IR scenarios, as not all terms are significant in respect of a text's underlying semantics. Commonly applied strategies are:
   - Filtering terms that occur often (i.e. stop words). In English, these include terms that occur in a majority of texts, e.g. *the*, *is*, *a* or *an*. By being present in many different articles, they are not significant for a certain topic or concept and thus do not contribute to a semantic differentiation.
   - Filtering terms that are very rare. In English, these commonly include terms that are very specific and technical terms (e.g. *grandiloquent* or *ultracrepidarian*), proper names (e.g. the Indian name *Prandharath*) and misspellings (e.g. *missspeling*).
   - Stemming [123, 151], which is subsuming terms that stem from the same root with their *stem*. For example, the terms *connected*, *connection* and *connections* share the common stem *connect*. By reducing all occurrences of the different forms to the stem, the number of terms is decreased without diminishing the semantic content of an article too much. In fact, accuracy in European languages can be improved significantly by working on stemmed indexes in retrieval settings [93]. However, stemming introduces confusion of lexically similar terms, the German Porter stemmer e.g. stems both German terms "Leber" (liver) and "Leben" (life) to the same stem "Leb". Using ESA, these both concepts can be discriminated only by their context, i.e. the surrounding terms.

- Filtering terms by their function as part of speech. The idea behind this strategy is that not all term forms equally contribute to the semantic content of a text. For example, as nouns represent the actors and objects of a sentence, they can be regarded to be more important than verbs that describe the action that is performed. This can be observed e.g. in tagging systems, where an overwhelming majority of tags consists of nouns or noun groups [81], thus users seem to express the semantics of texts primarily in nouns.

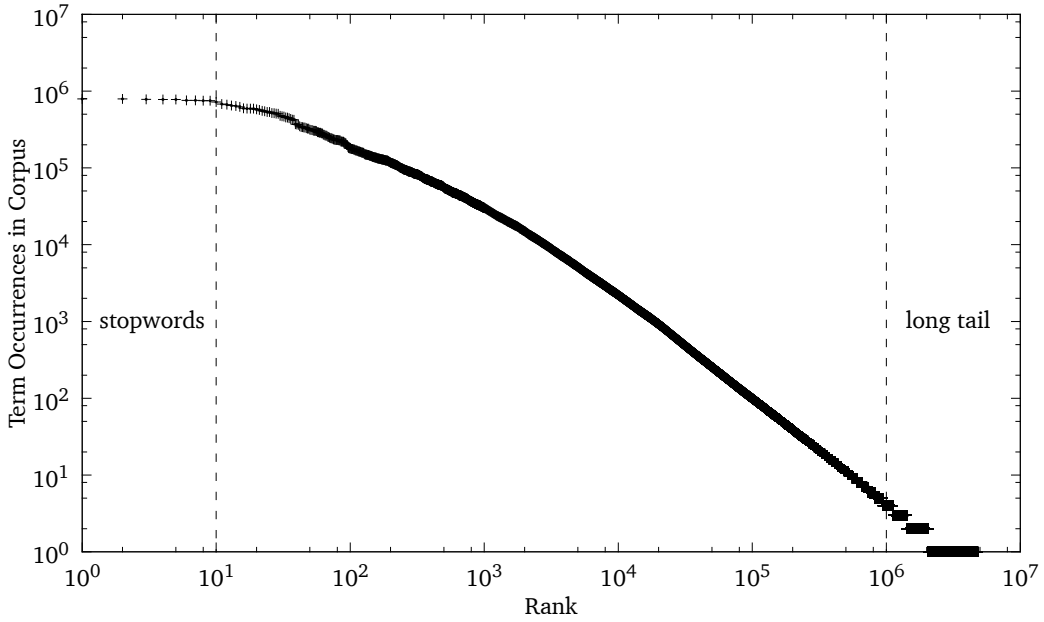


Figure 3.5: The ranked occurrences of terms of Wikipedia against the number of articles they appear in. Note that this curve shows only an approximate Zipfian distribution, because the term distribution is skewed due to stemming the terms. Further, the thresholds for stop words and long tail are only exemplary.

However, a reduction of the semantic interpreter may influence the results of the semantic analysis. On the one hand, the semantic analysis may benefit from the reduction, because irrelevant terminology and non–concept articles are filtered. This may greatly reduce semantic noise and increase the accuracy of ESA. On the other hand, if relevant terminology or articles that have describe a relevant concept are filtered, this may reduce the *global accuracy* of this approach, as the *coverage* (the existence of the source document's terminology and key concepts in the reference corpus, cf. section 3.3.2) is not ensured.

Further, an appropriate term or article reduction strategy always needs to take the target scenario into account. For example, a generic application scenario needs to have broad, general concepts in its semantic interpreter, whereas an application scenario that works with a specific, rare terminology should have a large terminology coverage. Thus, it is important to take into account the particularities of the scenario before applying a reduction strategy.

In the original ESA approach [79], Gabrilovich and Markovitch observe that not all articles within Wikipedia are equally significant in respect to their semantic content. They identify following properties of articles that make them less useful for calculating semantic relatedness:

Too short articles Gabrilovich and Markovitch remove articles containing less than 100 terms from the semantic interpreter.

Overly specific articles describe a concept that is very specific, even for a narrow domain of knowledge. More general concepts should be sufficient to represent a topic on their own, thus making overly

specific articles obsolete for the analysis. For example, the article *Xibalba*[18], the underworld of Mayan mythology, represents a concept that may be of interest for semantic relatedness in the special domain of Mesoamerican Archeology, but will rarely contribute to a representation of general knowledge. Or, another example is the article *Zebra Finch*[19] that can be omitted if enough information is already covered in the more general article *Estrildid Finch*[20]. Gabrilovich and Markovitch use the number of incoming and outgoing links as an indicator of its usefulness as a concept. As, according to the Wikipedia Guidelines[21], an article should link to another only when clarification or context is needed, this should result in more general articles having more incoming links than specific articles, because each article describing a specific topic will link to its generalized concept in order to clarify its context. In the original ESA approach, all articles having fewer than five incoming and five outgoing links are filtered.

Aggregate articles break the paradigm of "one concept — one article", as they represent a collection of links to other concepts with a certain categorization system. For example, *April 23*[22] is an article grouping events that occurred on this specific date and therefore is an example for a temporal classification system. These articles add unnecessary noise to the reference corpus because they do not contribute to a conceptual differentiation. Thus, in ESA, articles that describe specific dates or list events of a particular year are excluded from the semantic interpreter. Further, as disambiguation pages do not represent a single concept on their own but rather provide an itemization of different meanings of ambiguous terms (cf. [83]), these are left out due to the violation of the one–to–one concept–article mapping.

Gabrilovich and Markovitch eliminate these articles from the reference corpus, thus reducing the dimensions of the semantic interpreter matrix M. They state that this procedure reduces the 910,989 articles found in the Wikipedia snapshot of November 5, 2005 by approximately 81%, making the complete dataset processable and thus enhancing the performance of the analysis. This filtering step is claimed to produce better results, but this is not supported by an empirical evaluation. This means that there is no data on which effect changing these parameters have with respect to the accuracy of computing the semantic relatedness.

Further, studies by Zesch and Gurevych [203] on the effect of the growth of Wikipedia with respect to accuracy of ESA indicate that it performs equally well with older versions of Wikipedia, although these typically contain only a subset of the articles of the current version. The studies indicate that the accuracy of ESA increases proportionally to the number of articles in Wikipedia until a critical mass is reached. At a point where around 200,000 articles are used to build the semantic interpreter, the accuracy seems to stagnate.

In the following, several filtering strategies are evaluated concerning the coverage, accuracy and size of the respective semantic interpreters. The goal of this research is to reduce the semantic interpreter M while retaining the quality of ESA. Further, applicable parametrizations for ESA with respect to different scenarios are presented. It should be noted that all semantic interpreters used below contain articles that have been stemmed (unless stated otherwise), as the morphology of a term is not important for *calculating* semantic relatedness (although, it may be important for a human rater). In fact, *not* applying stemming can seriously impede such an approach, as ESA would for example not be able to map nouns

[18] `http://en.wikipedia.org/wiki/Xibalba`, *retrieved 2011-02-01*
[19] `http://en.wikipedia.org/wiki/Zebra_Finch`, *retrieved 2011-02-01*
[20] `http://en.wikipedia.org/wiki/Estrildid_finch`, *retrieved 2011-02-01*
[21] `http://de.wikipedia.org/wiki/Wikipedia:Artikel`, *retrieved 2011-02-11*
[22] `http://en.wikipedia.org/wiki/April_23`, *retrieved 2011-02-01*

in singular and plural forms, making it unable to identify *cats* and *cat* as the same term. Thus, the used semantic interpreters are already reduced by stemming.

3.4.1 Evaluation of Article Filter Strategy based on Link Type Selection

Gabrilovich and Markovitch [79] filter all articles that have less than five incoming links and less than five outgoing links based on the assumption that the remaining articles are general enough and rich enough in content for representing a semantic concept. However, there are alternative strategies based on link type selection that can be applied. For this evaluation, following strategies have been applied and evaluated:

- Filtering articles with less than a certain amount of incoming links (*inlinks*)
- Filtering articles with less than a certain amount of outgoing links (*outlinks*)
- Filtering articles with less than a certain amount of in– *and* outgoing links. This strategy, used by Gabrilovich and Markovitch, is merely an intersection of the article set produced by the first two strategies.
- Filtering articles with less than a certain amount of *mutual links*. A mutual links exists if an article a_1 links to article a_2 and vice versa.

Figure 3.6 displays the effect of these filtering strategies on the number of articles in the semantic interpreter. Note that disambiguation pages are not included in the articles used to generate the semantic interpreters.

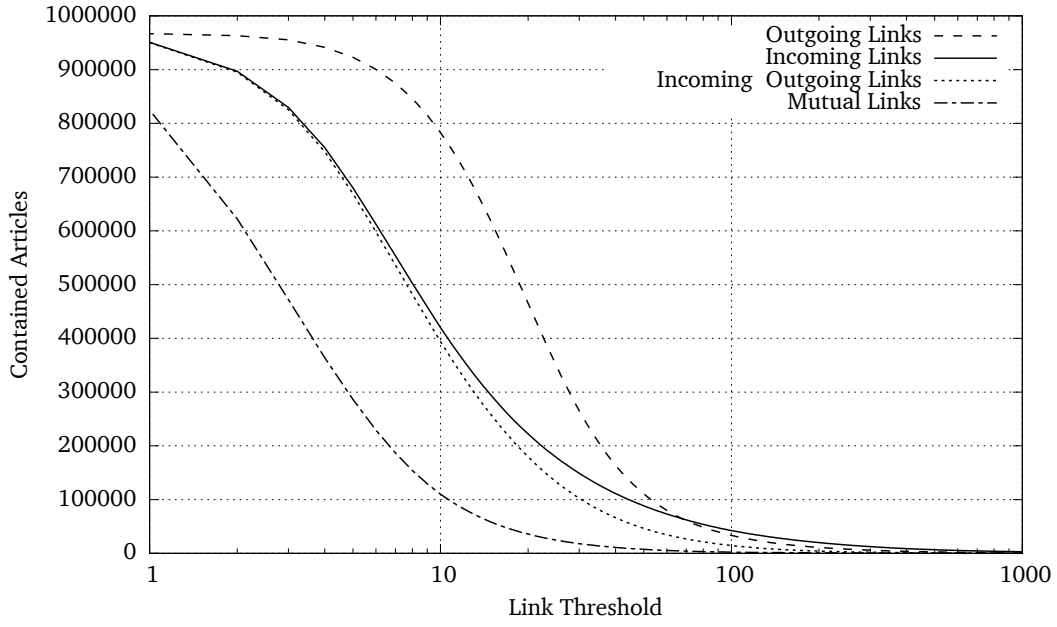


Figure 3.6: Number of articles depending on filtering strategies based on article linkage

This figure shows that filtering the articles by the number of different link types significantly decreases the number of remaining articles. Furthermore, the different strategies have a varying influence regarding the degree of decrease. Articles featuring many outgoing links are more numerous than articles having many incoming links. Curiously, this trend is reversed at the threshold of approximately 60 links. This is probably due to a core of very generic articles that are linked to very frequently (e.g. the article

England[23] that is linked to 16,172 times and only links to 92 other articles). The conjunction of inlinks and outlinks having the same threshold that is applied in ESA's original research is approximately approaching the curve of the inlink strategy. It is not explored in more detail in the following, because the semantic interpreters of the inlink strategy and the combined strategy are very similar. Further, there are less occurrences of articles containing many mutual links than using other strategies and accordingly the respective curve drops quickly. Already with a filter threshold of three, the number of remaining articles is only about 50%.

In the following, an evaluation is presented that analyses the accuracy benefits in relation to the size of the respective semantic interpreter. The conjunction of inlinks and outlinks is not considered here, as it is very similar to the inlink strategy.

	Inlink Filter		Outlink Filter		Mutual Link Filter	
Filter Threshold	Articles	SI non–zeros	Articles	SI non–zeros	Articles	SI non–zeros
0	973,227	197,640,359 (100.00%)	973,227	197,640,359 (100.00%)	973,227	197,640,359 (100.00%)
5	-	-	-	-	291,831	66,122,694 (33.46%)
10	471,504	119,482,379 (60.45%)	834,394	134,374,296 (67.99%)	113,018	35,174,062 (17.80%)
25	261,569	75,559,590 (38.23%)	401,434	122,292,448 (61.88%)	25,158	12,273,530 (6.21%)
50	149,370	48,703,205 (24.64%)	164,039	66,224,493 (33.51%)	-	-
75	101,082	35,996,128 (18.21%)	88,048	42,704,594 (21.61%)	-	-
100	74,270	28,248,574 (14.29%)	52,391	29,965,737 (15.16%)	-	-
200	39,406	15,966,763 (8.08%)	12,042	11,341,855 (5.74%)	-	-
300	31,452	11,983,460 (6.06%)	-	-	-	-
400	28,078	9,831,884 (4.95%)	-	-	-	-
500	26,436	8,654,881 (4.38%)	-	-	-	-

Table 3.6: Impact of filtering by inlinks, outlinks and mutual links on article count and semantic interpreter (SI) size

Table 3.6 shows the ratios of decrease of article count and semantic interpreter size (in non–zero matrix entries) on using inlink, outlink and mutual link filter strategies. Further, it indicates that the decrease of the corpus size for both inlink and outlink strategies is correlating to the decrease of the number of articles remaining in the corpus, although it is not proportional (cf. figure A.1 in appendix A.3). Mutual links are rare in Wikipedia, thus the number of articles drops quickly with already a low filter threshold.

The results of Zesch and Gurevych [203] indicate that the optimal number of articles taken into account for a semantic interpreter is about 200,000, however, this number is based on a complete Wikipedia dump containing all articles. It is expected that this number can be reduced by an appropriate article filtering strategy.

In order to get a conclusive picture about the quality of the link filtering strategies, the resulting semantic interpreters are compared with regard to following performance indicators (cf. section 3.3.2):

Coverage, Global and Local Accuracy using the Reader's Digest Word Puzzle Corpus `RDWP984`.

Correlation with human judgement based on Spearman's Rank Correlation Coefficient ρ using the two term–pair datasets `Gur65` and `Gur350`.

MAP and BEP based on experiments performed with the semantic corpus `Gr282`.

[23] `http://en.wikipedia.org/wiki/England`, *retrieved 2011-01-30*

Coverage and Accuracy

A reduction of articles is accompanied with the removal of specific terms that are relevant to the filtered articles. Thus, the coverage will decrease with an increasing level of filtering. Figure 3.7 shows the increase of coverage for the inlink, outlink and mutual link filter strategies with semantic interpreters containing more articles for different points of measurement. The points of measurement represent the different filtering thresholds, but they are transcribed to the number of articles that are contained in a semantic interpreter to make the results comparable. The outlink filter strategy approximately subsumes the mutual link filter strategy. Both perform better for semantic interpreters where less than approximately 250,000 articles are retained. For semantic interpreters containing more than this number of articles, the coverage is similar in both outlink and inlink strategies. A reason for this better performance of the outlink filter strategy could be that the number of outlinks in an article correlates with the size of the article in terms. Consequentially there is a bigger chance that long articles contain a larger diversity of terminology than short articles. Thus, the coverage of the outlink and mutual link filter strategies is better especially with smaller semantic interpreters.

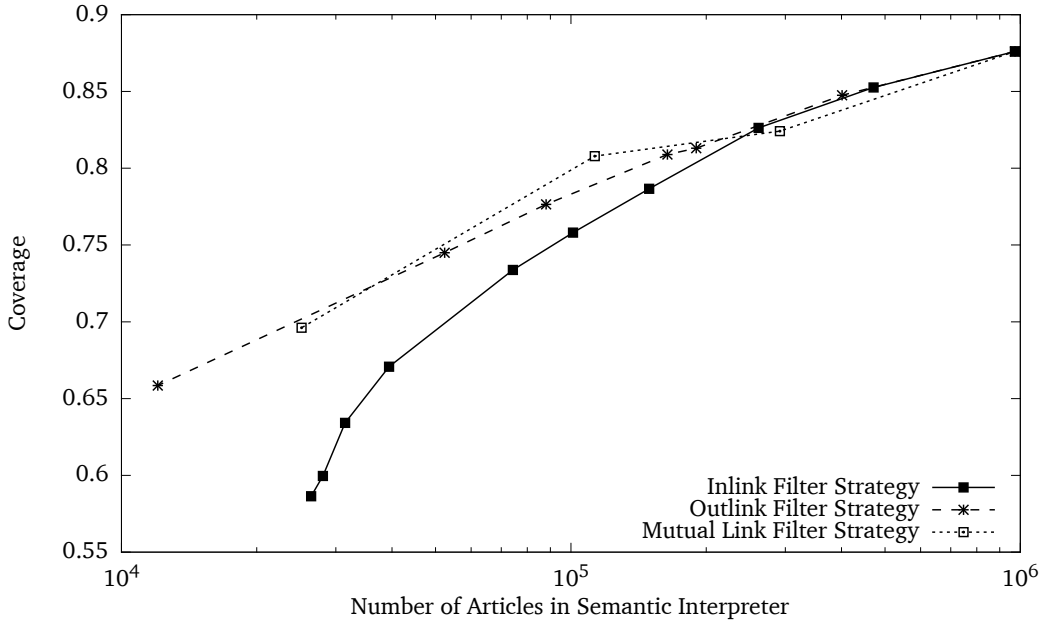
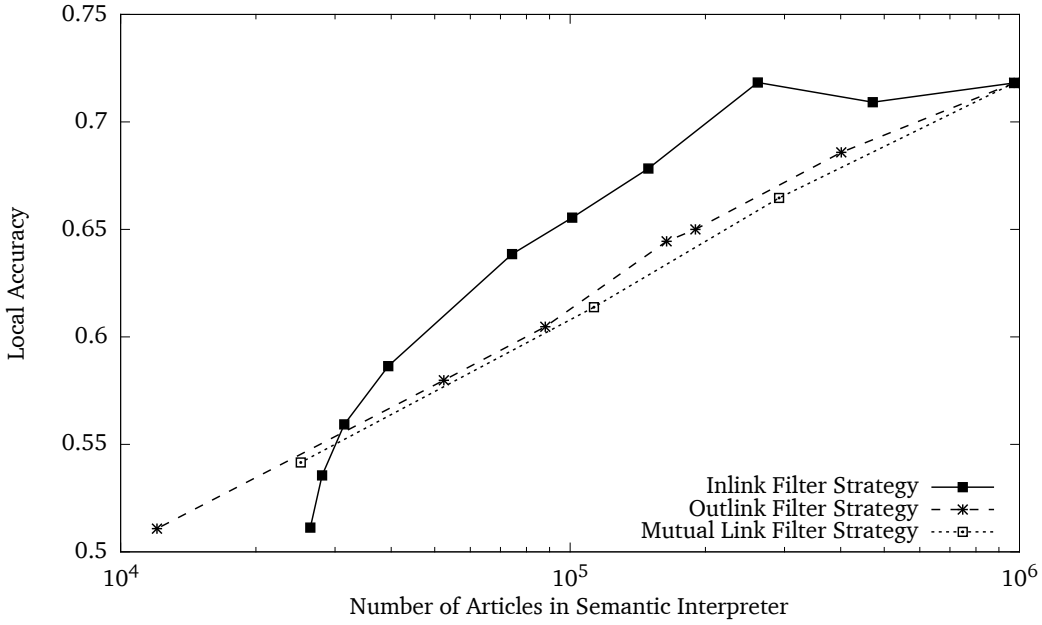


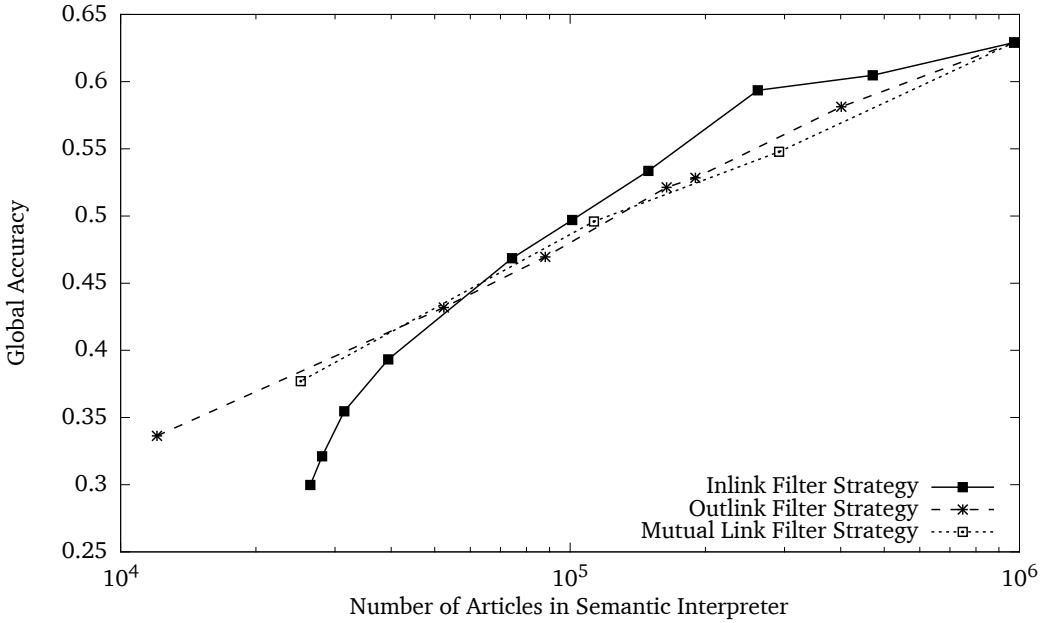
Figure 3.7: Coverage for inlink, outlink and mutual link filter strategies

According to [203], accuracy should be stable for a semantic interpreter consisting of about 200,000 articles. Figures 3.8a and 3.8b show that this not only applies to semantic interpreters containing all articles but also for the inlink filter strategy. A plateau of local accuracy is reached with approximately 220,000 articles for both local and global accuracy. The outlink filter strategy again subsumes the mutual link strategy. Both strategies, however, steadily increase the local accuracy until all articles are contained. For all semantic interpreters containing more than 30,000 articles, the outlink and mutual link filter strategies are dominated by the inlink filter strategy regarding the local accuracy. This is in accordance with global accuracy, which can also be seen in tables A.8 and A.9 in appendix A.

These results show that the accuracy of the inlink filter strategy performs better than the outlink and mutual link filter strategies for semantic interpreters that contain more than about 30,000 articles. This



(a) Local Accuracy



(b) Global Accuracy

Figure 3.8: Local and global accuracies for the Reader's Digest Word Choice Puzzle corpus `RDWP984` using inlink, outlink and mutual link filter strategies

supports the assumption of [79] that articles containing many inlinks are more generic and thus more applicable to be used as concepts in ESA.

So far, these experiments show that an increasing number of articles (and therefore concepts) used in the semantic interpreter primarily causes an increase of coverage while local accuracy is only affected to a small extent. Thus, more articles mean rarely used words can be mapped to concepts and compared to other terms contained in the semantic interpreter. This is crucial for the evaluation of semantic relatedness and semantic similarity due to the very limited number of available terms for the analysis.

Correlation with Human Judgements

In order to get a better understanding of the effect of filtering articles on ESA's ability to judge the semantic relatedness between terms, a second evaluation was executed. The German datasets `Gur65` and `Gur350` (cf. section 3.3.2) are used to measure the correlation between human judgements and computed judgements on semantically related word pairs. The results for both corpora can be seen in figures 3.9a and 3.9b.

These two evaluations show basically a similar tendency to correlate with semantic relatedness judgements of human raters as accuracy in the previous experiments. However, the inlink filter strategy visibly outperforms the outlink strategy above the range of about 30,000 contained articles. This is no surprise, as a similar characteristic could already be seen with accuracy. The mutual link filter strategy is clearly outperformed by both inlink and outlink filter strategies. Although it yields good results regarding coverage, it seems to eliminate important articles from the semantic interpreter M. Therefore, the mutual link filter strategy is not applied in the following experiments.

The steps between the points of measurement in figure 3.9a where large "jumps" in correlation can be seen for the inlink and outlink filter strategies are due to the increased coverage of the terminology. This affects the `Gur65` corpus more than the `Gur350` corpus, as it contains less word pairs. Interestingly, the inlink filter strategy in the `Gur350` dataset shows to have a local level of saturation at about 200,000. Here, the increase of the number of articles does not increase the correlation with the human judgements equally. The results for all points of measurement can also be found in tables A.8 and A.9 of appendix A.

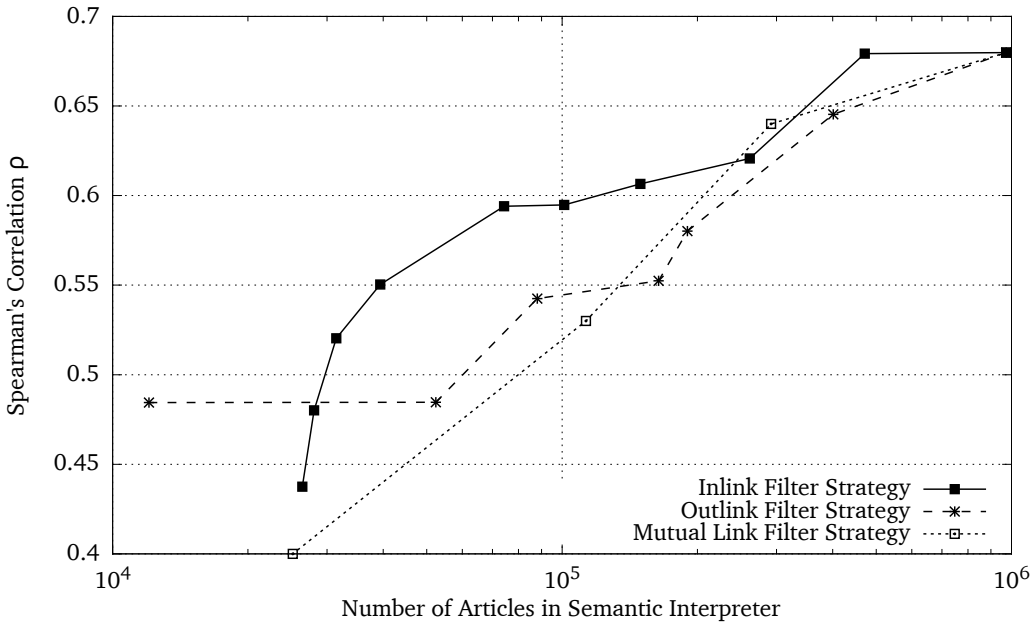
The findings up to now support the hypothesis that the inlink filter strategy is superior to the other strategies as long as the semantic interpreter does not include less than 30,000 articles. Thus, in the following evaluation only the inlink filter strategy is applied.

Detecting Semantically Related Documents

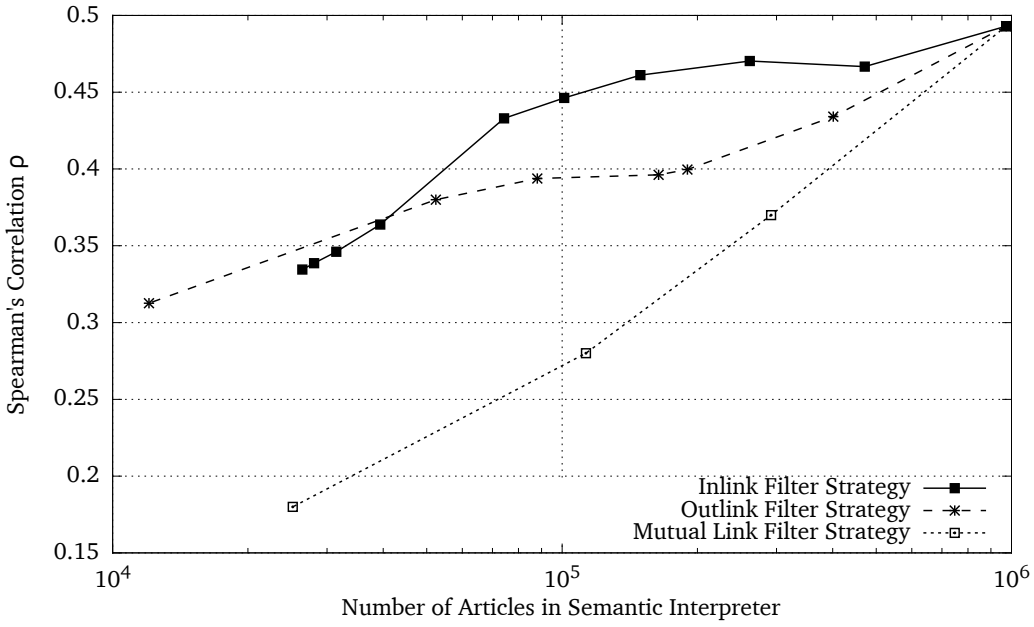
In most cases, documents of the above–mentioned corpora only consist of single terms. The `Gr282` corpus, however, is assembled from documents that contains 95 terms on average. As there are multiple terms, the effect on not covered terminology on the accuracy of a semantic interpreter is not as severe — even if some terms are not covered by the reduced semantic interpreter, the remaining terms can still be analysed. Therefore, reduced semantic interpreters could yield accuracy comparable to the accuracy of a semantic interpreter built from all available articles. In order to measure the impact of using reduced semantic interpreters in IR tasks, the following experiments are performed on the `Gr282` dataset.

The inlink filter strategy shows to be better suited to reasonably reduce the semantic interpreter. Therefore, only this strategy is applied in the following experiment. The effects of different link filtering thresholds are shown in figure 3.10 and table A.10 in appendix A.2.

The BEP and MAP values in figure 3.10 show that the performance of ESA increases up to the point of measurement with a minimum of 100 incoming links. Subsequent measurements with a larger amount of



(a) Corpus `Gur65`



(b) Corpus `Gur350`

Figure 3.9: Comparison of the effect of inlink, outlink and mutual link filter strategies and different sized semantic interpreters on the correlation between human and computed relatedness judgements.

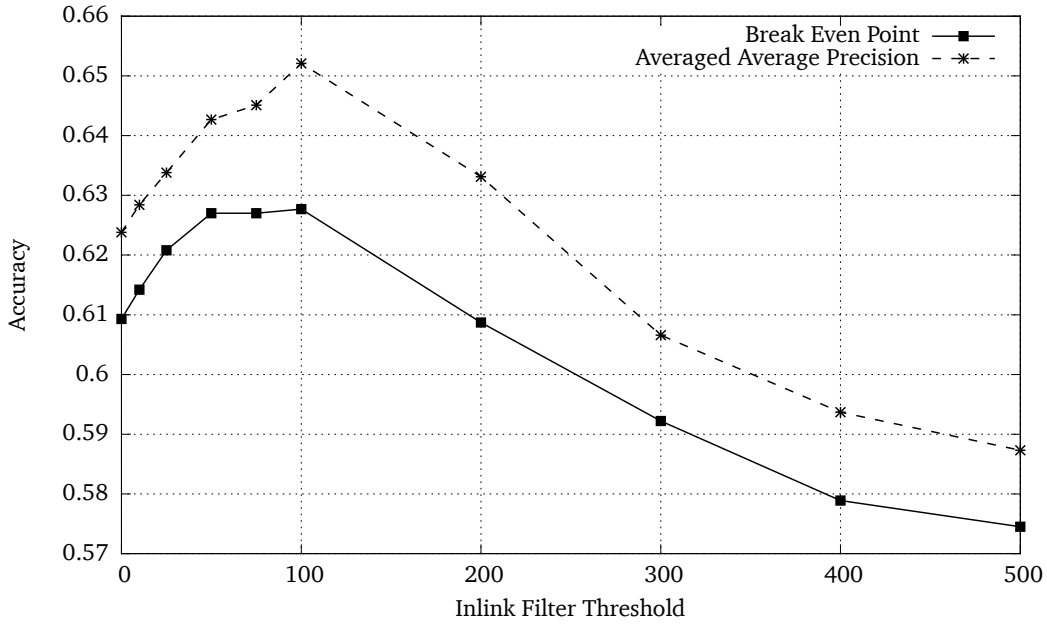


Figure 3.10: The Mean Average Precision and Break Even Point of dataset `Gr282` using semantic interpreters reduced by the inlink filter strategy.

articles do not show a gain in terms of accuracy. In fact, it already begins to decrease slightly at small link filtering thresholds. An explanation for this observation is that noise is introduced by articles that do not cover the most relevant terms but still have a certain term overlap. Therefore, the terminology coverage of the used corpora is already near–complete and additional articles do not significantly contribute to the information contained in the semantic interpreter but rather distort the semantic analysis. Another explanation for this observation is that short articles do not contain sufficient terms that appropriately describe the article's concept.

3.4.2 Evaluation of Article Filter Strategy based on Heuristics

As shown in section 3.4, there are types of articles that violate the paradigm that one article should describe exactly one specific concept. This has implications on the applicability of a semantic interpreter: for example, disambiguation pages serve to discern between homonyms, presenting each concept with a short description and a link to the respective article. Often, these pages are highly rated in the semantic analysis, as they contain the important terminology for a concept in a short document. Thus, they are frequently assessed to be more relevant to a given document than the "real" article describing exactly the concept that is covered in the document. Thus, the existence of disambiguation pages in a semantic interpreter decreases its ability to disambiguate between homonyms.

Disambiguation pages are not the only page types that generate noise in the semantic interpreter. In generic settings that need calculation of semantic relatedness, different article types can be filtered in order to reduce the semantic interpreter and eliminate noise. However, the scenario has to be analysed before filtering possibly relevant article types. For example, a hypothetical scenario that needs to compute the semantic relatedness between rock albums should not filter Person articles, as they contain the songwriters and artists as concepts.

The following article types are considered as being candidates for filtering as well:

Event articles are an aggregation of events that happened during a certain time span and therefore introduce a temporal association of concepts. Gabrilovich and Markovitch [79] apply this filter heuristic, as they state that Wikipedia articles like *April 23* serve as a collection of events that happened at the same day but do not describe a single underlying concept. A similar article type lists important events of certain years, decades, centuries or millennia (e.g. *1984*, *1980s*, *20th century* or *1st millennium BC*). These articles introduce links between events that rarely have an impact on their semantic relatedness and therefore can usually be filtered.

Person articles describe the life and the achievements of famous or noteworthy persons. Although they adhere to the paradigm that articles should describe exactly one concept, they are often not needed in a scenario which encompasses general knowledge about a domain. Further, person articles introduce semantic connections between different concepts that are not closely related, e.g. the article about Albert Einstein[24] contains the terms *Ulm*, *Genie*, *Zionism* and *Citizenship*. Thus, such articles contain terms of a wide range of semantic concepts and therefore may introduce noise.

Listing articles provide collections of links to different articles based on certain criteria. For example, there are lists that collect articles about artificial intelligence projects[25] or Byzantine emperors[26].

Category articles serve to provide a hierarchical, vertical structure to Wikipedia by grouping articles and other categories. Usually, they contain no or little textual information and therefore do not provide a description of a concept. Often there are articles that describe the semantics of a category better than the category itself does (e.g. "Category:Horses"[27] groups different sub–categories and articles that are related to the concept "Horse", whereas the concept itself is described in the article "Horse"[28]). Therefore, categories can usually be filtered without loss of distinctive concepts.

Portal articles provide an entry point for members of a special interest group. They give a short overview of a certain field of interest, linking to the relevant articles that belong to this field. Therefore, even if portal articles contain textual information, the concepts they describe are covered in their respective articles in more detail. In Wikipedia, Category and Portal articles are treated differently by design than concept articles and they even have an own namespace with their lemma being prefixed with "Category:" respectively "Portal:". In the following examinations, they are therefore both treated in one single filter.

This listing is not exhaustive, arguably, there are further article types that can be filtered in certain usage scenarios. However, the structure of Wikipedia often does not specifically distinguish these types, an automatic classification is not feasible or the article types are insignificantly rare (e.g. articles that describe English language soundtracks[29] and therefore are too specific to describe a reasonable concept). In most approaches that build on using only the article contents of Wikipedia, at least some of the above mentioned article classes are filtered. To the best knowledge of the author, the Person and Listing filters are novel.

Table 3.7 shows how many articles are affected by each of the above–mentioned heuristics. It shows that a large fraction of all Wikipedia articles belong to these classes, 43.37% of all articles can be filtered by applying all heuristic reduction strategies. The most impact on the size of the semantic interpreter is achieved with the Person Filter. Indeed, the German Wikipedia contains a large fraction ($\approx$ 30%)

[24] `http://en.wikipedia.org/wiki/Albert_Einstein`, *retrieved 2011-03-01*
[25] `http://en.wikipedia.org/wiki/List_of_artificial_intelligence_projects`, *retrieved 2011-03-01*
[26] `http://en.wikipedia.org/wiki/List_of_Byzantine_emperors`, *retrieved 2011-03-01*
[27] `http://en.wikipedia.org/wiki/Category:Horses`, *retrieved 2011-03-01*
[28] `http://en.wikipedia.org/wiki/Horse`, *retrieved 2011-03-01*
[29] `http://en.wikipedia.org/wiki/Category:English-language_soundtracks`, *retrieved 2011-03-20*

Applied Filters	Remaining Articles (%)	SI nonzeros (%)
None (all articles)	1,095,678 (100.00%)	153,096,009 (100.00%)
Disambiguation Filter (DA)	973,227 (88.82%)	148,957,078 (97.29%)
Category/Portal Filter (CP)	1,085,519 (99.07%)	151,799,443 (99.15%)
Person Filter (PA)	770,726 (70.34%)	110,916,646 (72.45%)
Listing Filter (LF)	1,075,206 (98.13%)	148,039,424 (96.69%)
Event Filter (DF)	1,094,169 (99.86%)	152,165,780 (99.39%)
All (CP-DA-DF-LF-PA)	620,540 (56.63%)	97,276,300 (63.53%)

Table 3.7: Impact of different article filtering heuristics on semantic interpreter size in article size and non–zeros

of articles describing persons. The other filters (with the exception of the Disambiguation Filter) only marginally reduce the semantic interpreter.

	RDWP984					Gur65	Gur350
Applied Filters	Covered	Wrong	Correct	Global Accuracy	Local Accuracy	Correlation ρ	Correlation ρ
None (all articles)	911	244	667	68.78%	73.22%	0.69	0.51
Disambiguation Filter (DA)	910	246	664	67.48%	72.97%	0.65	0.49
Category/Portal Filter (CP)	910	245	665	67.58%	73.08%	0.69	0.51
Person Filter (PA)	906	232	674	68.50%	74.29%	0.74	0.55
Listing Filter (LF)	909	241	668	67.89%	73.49%	0.69	0.51
Event Filter (DF)	911	243	668	67.89%	73.33%	0.69	0.51
All (CP-DA-DF-LF-PA)	898	233	665	67.58%	74.05%	0.74	0.50

Table 3.8: Results of different article reduction strategies based on heuristics.

As table 3.8 shows, even on filtering these article classes, the coverage and quality of the evaluations does not decrease considerably. For the filters that only marginally reduce the semantic interpreter, this result is expected, as the relevant concepts are not affected. In the case of the `Gur65` dataset, the Person Filter even increases the correlation by removing irrelevant articles that introduced semantic noise into the semantic interpreter. It can be assumed that this is mainly due to activation of concepts that are closely linked to the respective terms. Overall, the semantic interpreter M is reduced significantly by these filter strategies to 63% of its original size without impairing the semantic relatedness results considerably. Therefore, the hypothesis that these article classes do not contribute to semantic relatedness calculation and therefore can be removed holds. Further, due to the filtering of person articles that associate possibly unrelated concepts, the quality of the calculation of semantic relatedness of terms can be increased.

3.4.3 Evaluation of Filtering Rare Terms

As figure 3.5 shows, a large number of terms is only present in a few articles. Some reasons for this are misspellings, terms borrowed from foreign languages (e.g. Arabian name of a person like *al–Chwarizmi*) and uncommon word compounds (e.g. *Nasenbeutlerspuren* which is *tracks of a bandicoot* in English). These long word compounds specifically occur in Germanic languages.

Terms which are very uncommon can be removed from a semantic interpreter in order to reduce its size. The results of such a reduced semantic interpreter are unlikely to differ from results obtained from a semantic interpreter containing all terms, as long as no term covered in an analysed document

is removed. Especially for analysing longer documents, the impact should be barely noticeable because they typically contain several terms that hint to the underlying concept. Thus, even when one rare term is not covered by the semantic interpreter, the remaining terms should be able to contribute enough information to analyse the document.

Threshold	Number of Terms (%)	SI nonzeros (%)	Gur65 (ρ)	Gur350 (ρ)	Gr282 (BEP)	Gr282 (MAP)
0	1,938,969 (100.00)	28,248,574 (100.00)	0.59	0.43	0.6277	0.6521
1	743,745 (38.36)	27,053,944 (95.77)	0.59	0.43	0.6250	0.6450
2	489,261 (25.23)	26,544,976 (93.97)	0.59	0.42	0.6204	0.6428
5	306,423 (15.80)	25,928,681 (91.78)	0.59	0.41	0.6180	0.6394
10	168,557 (8.69)	25,036,287 (88.63)	0.57	0.38	0.6174	0.6381
25	79,170 (4.08)	23,700,660 (83.90)	0.58	0.35	0.6172	0.6384

Table 3.9: Effect of filtering rare terms from a semantic interpreter on accuracy measures of different corpora. The threshold denotes the minimum occurrence of a term in order to be included in the semantic interpreter.

In order to achieve results which are comparable to the other experiments, a semantic interpreter built from articles with at least 100 inlinks was used. The results (cf. table 3.9) show that with filtering rare terms, the dimensionality of the term dimension of the semantic interpreter decreases considerably, whereas the number of its nonzero–entries is reduced less significantly. In the Gur65 dataset, the correlation is degrading step–wise with a higher term filtering threshold as the coverage decreases. As this dataset contains more general terms, the impact of filtering is less severe than in the Gur350 dataset, where rare terms occur. In contrast to the single–term corpora, the Gr282 dataset is not affected that considerably, because it contains multi–term documents.

3.4.4 Evaluation of Filtering Stop Words

Common terms like *the*, *and* or different forms of *to be* are used in a majority of the articles in Wikipedia. Thus, they do not contribute to discriminating between different concepts. Although ESA already utilizes the *tf–idf* measure to account for this irrelevancy of terms, storing them in a semantic interpreter takes up a lot of space without any benefits. Further, stop word removal removes densely populated columns from the semantic interpreter, thus decreasing its size considerably in terms of nonzero entries (see table 3.10).

Removed Stop Words	SI non–zeros (%)	Gur65 (ρ)	Gur350 (ρ)	Gr282 (BEP)	Gr282 (MAP)
0	28,248,574 (100.00%)	0.58	0.43	0.6277	0.6521
1	28,176,603 (99.75%)	0.58	0.43	0.6300	0.6494
5	27,902,520 (98.77%)	0.58	0.43	0.6331	0.6533
25	26,927,855 (95.32%)	0.58	0.43	0.6443	0.6733
50	26,136,435 (92.52%)	0.58	0.41	0.6717	0.7079
75	25,532,186 (90.38%)	0.58	0.37	0.6844	0.7261
100	25,037,242 (88.63%)	0.58	0.37	0.6944	0.7357
150	24,221,173 (85.72%)	0.58	0.35	0.7094	0.7526
500	20,941,401 (74.13%)	0.55	0.25	0.7417	0.7896

Table 3.10: Effect of stop word filtering on size of semantic interpreter in non–zeros and on the Gur65, Gur350 and Gr282 datasets.

For calculating the semantic relatedness of term pairings and very short snippets, removing stop words only has an effect if the snippets' terminology is removed from the semantic interpreter. This is reflected by the step–wise decrease of the correlation values for the Gur65 and Gur350 datasets. However, for longer documents, these common stop words introduce noise, as even many small values (like they are generated by *tf–idf* for common terms) accrue to a certain level of relatedness. The results of experiments using the Gr282 corpus show that the performance of ESA is increased for an incrementing number of removed stop words. This is because stop words introduce a deceptive relatedness, and after filtering, only terms remain that have a higher probability of describing a specific concept. Thus, especially in such a document–level setting, removing stop words has a beneficial impact on the quality of ESA.

3.4.5 Evaluation of Filtering based on part–of–speech tags

In most languages, the fundamental parts of a sentence are subjects and verbs, often in conjunction with an object. Other parts of speech modify or enrich these basic parts. For example, adjectives may be used to modify a noun or a pronoun. Adverbs may be used to modify a verb, adjective or another adverb. As described above, nouns or noun groups are mostly used as tags for describing content in tagging systems. Thus, for ESA, the following experiment should determine, whether the basic semantics encoded in nouns or other parts of speech are sufficient for representing the concepts contained in Wikipedia.

For determining the parts of speech, the part–of–speech (POS) tagger TreeTagger[30] [171] is used to tag all articles of Wikipedia. Then, only the terms that belong to the inspected POS groups are taken into account for building the semantic interpreter. Because the tagging process is computationally expensive, articles with fewer than 200 inlinks are ignored in this experiment in order to reduce the complexity. Four different variants of POS selection are applied:

- All parts of speech (e.g. nouns, verbs, adjectives and others, in their normalized form as returned by TreeTagger).
- Nouns only (tagged by TreeTagger with the POS tags NN, NNS, NP and NPS, cf. the Stuttgart–Tübingen Tagset [170]).
- Verbs only (tagged with the POS tags VB, VBD, VBG, VBN, VBP and VBZ).
- Nouns and verbs (all of the above tags).

POS selection	Number of Terms (% Terms)	Nonzero entries in SI	BEP	MAP
All Parts of Speech	1,372,584 (100.00%)	15,966,763 (100.00%)	0.61	0.60
Nouns	1,212,583 (88.34%)	9,645,882 (60.41%)	0.73	0.75
Nouns ∪ Verbs	1,241,179 (90.43%)	11,585,094 (72.56%)	0.69	0.72
Verbs	38,029 (2.77%)	1,946,630 (12.19%)	0.18	0.14

Table 3.11: Reduction of semantic interpreters by part–of–speech selection. "All Parts of Speech" contains all terms in the corpus in their normalized form. The BEP and the MAP show the predominance of including nouns in a semantic interpreter M in contrast to verbs.

For the latter three experiments, all terms that are not identified as a noun or verb are discarded. Table 3.11 shows the resulting sizes of the semantic interpreters and their respective results on the Gr282 dataset. The largest part of the articles consists of nouns. This supports the hypothesis that nouns

[30] Available at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html, *retrieved 2011-02-09*, TreeTagger was used with the standard German parameter file.

are in fact an appropriate means to express semantics in natural language. Further, a slight difference between the number of terms which are nouns summed with the number of terms which are verbs and the number of terms which were identified as nouns or verb can be seen. To some extent, this difference can be caused by an erroneous POS tagging procedure. Further, a part of this overlap is generated by terms which can be used as nouns and as verbs as well (e.g. terms like *Leben* (life) or *leben* (to live) can be used as a verb or noun).

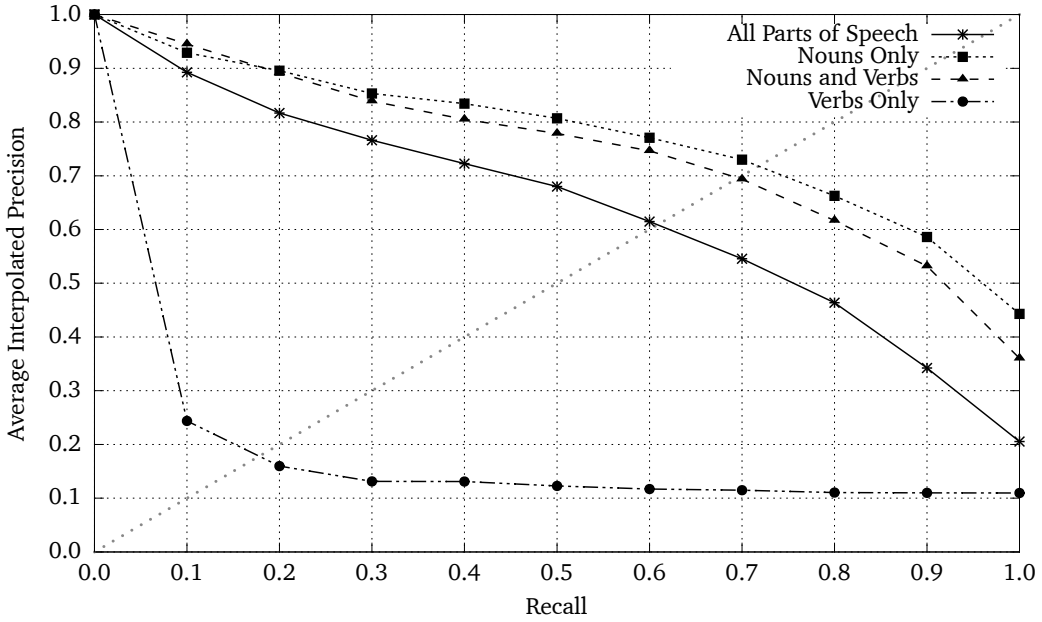


Figure 3.11: ESA used on `Gr282` with different parts of speech

Figure 3.11 shows the impact of POS reduction on the `Gr282` dataset. It shows that verbs do not capture the semantics of a text well and do in fact introduce noise to the semantic interpreter, whereas nouns are good descriptors for semantic concepts. In comparison to the semantic interpreter using all POS forms, the interpreter reduced to nouns shows an improvement in accuracy.

3.4.6 Reduction of the Semantic Interpretation Vector

Applying ESA results in the semantic interpretation vector i_{esa} that denotes the terminology similarity between the source document and the concepts represented in the semantic interpreter M. Documents that contain many common terms thus yield an interpretation vector that contains similarity values for many concepts. Therefore, the density of such an i_{esa} is high. Experiments have shown that for German multi–term documents, an interpretation vector often has up to 600,000 non–zero entries, of which a majority is usually negligibly small (< 0.001). This is due to the occurrence of terms that are frequently used but are not yet stop words. Further, in the comparison between two interpretation vectors, these values (albeit very small) introduce noise, decrease the impact of semantically relevant concepts and therefore contort the relatedness.

Further, in a setting that involves comparing a possibly large amount of interpretation vectors, the storage space requirements of such vectors have to be considered. For example, in ELWMS.KOM, where the semantic analysis is not performed on the fly but each new document is transformed to a interpreta-

tion vector once, these vectors have to be stored efficiently. Thus, if the dimensions of these vectors can be reduced, the needed space and the calculation time for vector comparisons decrease.

Similarly, in settings where the calculation of i_{esa} is an intermediate step (e.g. for multilingual semantic relatedness in section 3.5 or extended ESA in section 3.6), the performance of the approach can be enhanced significantly.

Therefore, in this thesis the function *selectBestN* is defined that reduces i_{esa} to its core containing only the n most relevant concepts, i.e. the concepts with the highest similarity values.

$$i_{esa}^{n} = selectBestN(i_{esa}, n) \qquad n \in \mathbb{N} \tag{3.10}$$

However, the parametrization of n is heavily dependent on the application scenario. For example, when measuring the semantic relatedness of two terms, the probability that relevant information is discarded is higher with a low n than when determining the relatedness between two medium–sized documents. Sorg and Cimiano [181, 54] report that using a higher n increases the recall for a specific concept. Therefore, in the following presentations of results, the impact of *selectBestN* is shown for respective evaluations.

3.4.7 Conclusions of Optimization Strategies

In this section, several strategies to reduce the concept and term space of the semantic interpreter have been presented and evaluated. Depending on the application scenario, these results allow to draw conclusions about an applicable parametrization of building the semantic interpreter. A good parametrization reduces the computational complexity and hard disk space requirements of ESA and — with some strategies — even lowers the impact of noise, resulting in a higher accuracy than the original ESA parametrization.

For most scenarios where the semantic relatedness should be calculated based on general concepts, article filtering strategies based on inlinks, outlinks and heuristics show to be promising. Especially the inlink filter strategy usually outperform the outlink filter strategy on semantic interpreters with more than 100,000 articles. Further, removing rare terms does only negatively impact settings where a very specific terminology has to be covered. Removing stop words positively affects accuracy mainly in settings which encompass the computation of semantic relatedness between multi–term documents, whereas coverage of single–term settings is only decreased if the terms are very generic.

In conclusion, there are two different use–cases in ELWMS.KOM that can be appropriately targeted:

Recommendation of Tags As tags are usually consisting of few terms, a parametrization can be derived from the results of the term–term relatedness experiments. Determining a good filtering strategy depends on a trade–off between the terminology that should be covered, the quality and the size of the semantic interpreters. For example, a scenario where tags are expected to be general, the coverage of rare terminology is not vital and thus, a rare term filter can be applied. However, the stop word threshold should not be set too high in order not to remove important generic terminology. Article filters have shown to not affect the quality of term–term relatedness ESA (with the notable exception of the Person article filter, which even increased the correlations for the used datasets), thus they should all be applied. The Person filter is especially useful as it reduces the semantic interpreter considerably, but it should only be used if Person concepts are not necessary for describing the used tags.

Recommendation of Snippets Snippets contain more textual context than tags, and thus should be handled with a different parametrization of the semantic interpreter. For the `Gr282` dataset, the results benefit most from the filtering of stop words. Filtering rare terms does decrease the results marginally, thus an application of both term filtering strategies should be considered. Further, if the individual snippets contain enough nouns, a POS filter can be applied.

For both scenarios, the inlink filter strategy has proven to have more stable results than the outlink and mutual link filter strategies, thus it should be applied as an additional filter. However, the appropriate filtering threshold differs on the number of articles that are contained. The number of articles that seems to fit best to all examined datasets is between 100,000 and 200,000 for the German Wikipedia (i.e. an inlink filter threshold between 50 and 100). However, this has to be examined and confirmed if ESA is to be used in other languages.

In the following evaluations, the used standard parametrization is comparable to the original ESA (unless specified otherwise) in order to allow comparison with the results of related work.

3.5 Cross–Language Relatedness using ESA

As shown in section 3.1.2, users of ELWMS.KOM often store resources in different languages. Therefore, an approach calculating semantic relatedness should provide a way to bridge the gap between languages appropriately. In this section, an approach to comparing semantically related terms and documents in different languages is presented.

As the examined scenario in ELWMS.KOM predominantly involves German and English LRs and tags, the following sections are confined to an analysis of those two languages. The general transferability of the proposed approaches to other language pairs is discussed in subsection 3.5.4.

3.5.1 Choice of Language Space and Transformation

Due to the good results of ESA in monolingual settings (cf. [79, 3, 203]), many researchers have applied it to cross–lingual contexts and their evaluations show promising results [152, 181, 86] (cf. section 3.2.4). As described by Cimiano et al. [54], this explicit approach also clearly outperforms any latent approach in cross–lingual systems.

However, the original ESA is an inherently language dependent approach, because the semantic interpretation matrix is dependent on terms from a specific language. In order to transform ESA to a multilingual approach, the different languages have to be mapped to a common language space at some stage of the relatedness calculation. Thus, for the design of an approach that provides cross–language semantic relatedness calculation, the fundamental decision has to be made how to design the language space and when to map it. Basically, there are three different possibilities:

- Translating all documents into one target language (similar to [137]) and apply monolingual ESA in this target language space. This is a naïve approach to handle the cross–lingual mapping that relays the challenge to the ability of the translation engine. Thus, its performance is heavily dependent on the quality of the employed translator.
- Creating a cross–lingual semantic interpreter and apply ESA to documents that have already been transformed to the cross–lingual hyperspace. Therefore, a language hyperspace is introduced that is not bound to a single language but rather to a set of languages. In accordance, each concept dimension of the semantic interpreter M is not addressed by the lemma of the corresponding

Wikipedia article in a single language but by the set of lemmata of articles describing the concept in the supported languages. The term dimension consists of sets of translations in the supported languages, e.g. produced by using a dictionary. For example, considering the set of languages $\{l_{en}, l_{de}, l_{fr}\} \in L$ (where L is the set of all languages that have an own Wikipedia), the English sentence fragment "the big house" is mapped to the bag of word triplets $\{\text{the}, \text{das}, \text{la}\}$ $\{\text{big}, \text{große}, \text{grande}\}$ $\{\text{house}, \text{Haus}, \text{maison}\}$. The result of this step is the language hyperspace L^* which is not a concrete language but rather a meta–language where each language element contains a set of elements from different concrete languages. However, translations are context sensitive and can rarely be mapped one–on–one, for example, the English definite article "the" needs to be added in several combinations, as it can be translated to "der", "die" and "das" in German and "le", "la" und "les" in French. The resulting term dimension would grow significantly with each added language. Further, due to the ambiguity of translation, a considerable amount of noise would be added. This approach, although hypothetically possible, is therefore barely practical.

- Applying ESA in each respective document's language space and map the resulting interpretation vectors to a common language using Wikipedia's interlanguage links. The common language can be either one of the languages of a document or a third language that serves as a unified reference language. This approach heavily depends on the quality and the amount of the interlanguage links.

In this thesis, the third possibility is chosen, because it is more practical than unifying all targeted languages in one hyper language space. Further, Semantic relatedness is defined on the assumption that humans' association of concepts provide a ground truth. Thus, the quality of the interlanguage links can be considered good, as they are manually set by humans. The quantity of interlanguage links is dependent on the choice of languages that are used in the specific setting. However, between German and English, the number of interlanguage links is considered sufficient [181, 54].

3.5.2 CL Links and Meta CL Links

The mapping process of semantic interpretation vectors is the crucial step in cross–language semantic relatedness calculation, as it determines the quality of the approach: the more concepts can be mapped from one language to the other, the closer the quality of interlingual semantic relatedness matches that of monolingual ESA.

Often, an interlanguage link (called CL link) exists for a Wikipedia article. This link represents the best possible mapping from one language space to another, because it interlinks two articles that ideally describe exactly the same concept. Thus, transferring the semantic interpretation vector i_{esa} from source language $l_s \in L$ into target language $l_t \in L$ just requires mapping the concept similarities to l_t using existing CL links. Thus, if a complete mapping exists between l_s and l_t, the cosine similarity between the interpretation vectors is 1.0 and therefore the quality of the cross–language ESA levels that of monolingual ESA. This approach has been described and evaluated by Sorg et al. [181], who show to achieve good results. If no CL link exists for an article, there might still be a corresponding article in the target language where simply the link is missing. Just applying a translation engine does not necessarily help here, as named entities often are not to be translated. Further, a term based translation adds lexical ambiguity. However, this can be dealt with by the approach presented in [177] (cf. section 3.2.4). Another issue is that not all existing CL links map articles one–to–one, but often l_s has an article a^{l_s} that encompasses a topic that is represented by multiple articles $a^{l_t}_{1..n}$ in l_t, which refer to their respective paragraphs in a^{l_s} covering the same concept.

Further, not all articles have CL links to all other languages that are represented by an own Wikipedia. For example, approximately 55% of German Wikipedia articles have a CL link to respective English articles, whereas only 18% of the English articles link back. Considering the size differences between the two Wikipedia versions, it can be assumed that nearly all CL links between those languages are bijective, i.e. $a^{l_{de}} \xrightarrow{\text{CL link}} a^{l_{en}}$ and $a^{l_{en}} \xrightarrow{\text{CL link}} a^{l_{de}}$. Missing CL links force to discard dimensions in the interpretation vector without a corresponding article in the target language. This issue can be dealt with by introducing subsidiary CL mappings between two articles in different languages that do not exactly describe the same concept but are somehow related to each other. There are several different approaches to infer the relatedness between missing concept translations in order to substitute CL links. Three of them are described in detail in the following paragraphs.

Usage of Chain Links

Chain Links [182] (cf. section 3.2.4) introduce links from articles not having direct CL links via articles that have a respective representation in the other language (see figure 3.12).

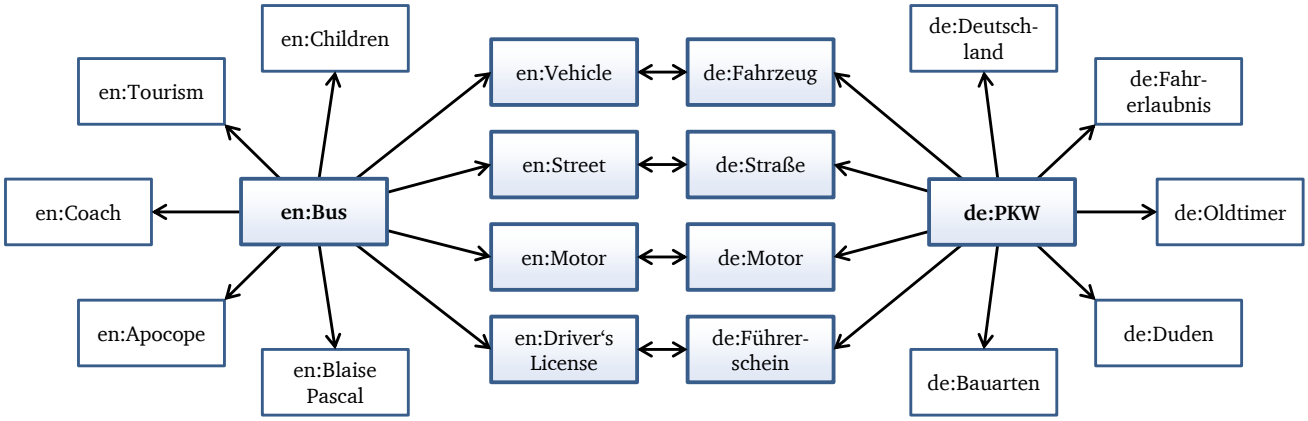


Figure 3.12: Example for Chain Links between the English concept *Bus* and the German concept *PKW*

However, as Wikipedia is densely linked, there are a lot of such chains to be found, even for unrelated or only marginally related articles. For example, the English article *Book* and the German article *Erbse* (pea) are connected via the following link structure:

$$\text{Book}^{l_{en}} \xrightarrow{\text{article link}} \text{Egypt}^{l_{en}} \xrightarrow{\text{CL link}} \text{Ägypten}^{l_{de}} \xleftarrow{\text{article link}} \text{Erbse}^{l_{de}} \tag{3.11}$$

In order to cope with those irrelevant Chain Links, Sorg and Cimiano introduce a lower threshold that the number of Chain Links between two articles has to exceed before two articles count as related. The number of Chain Links determines the degree of relatedness, thus the strength of the relation is assumed to be higher when more Chain Links exist.

From the candidates derived from the Chain Links, one is considered to be the "best match" for the article having no correspondent article in the target language. Sorg and Cimiano [182] use an approach based on classification in order to find the best match. However, in many cases there is no perfect match and the found candidate does not correspond to the concept described in the source article. In this case, the mapping generates wrong results, and this introduces a considerable amount of noise in the semantic interpretation vector i_{esa} eventually. Therefore, this approach is neglected in favour of another approach that reduces this source of error.

Usage of Categories

Categories group articles in specific topics. On average, a German article is categorized in 3.2 categories, whereas an English article is grouped in 3.6 categories. If two articles have a high ratio of overlapping categories, they can be considered to describe related concepts. This can be transferred to the inter–lingual space using category CL links (see figure 3.13).

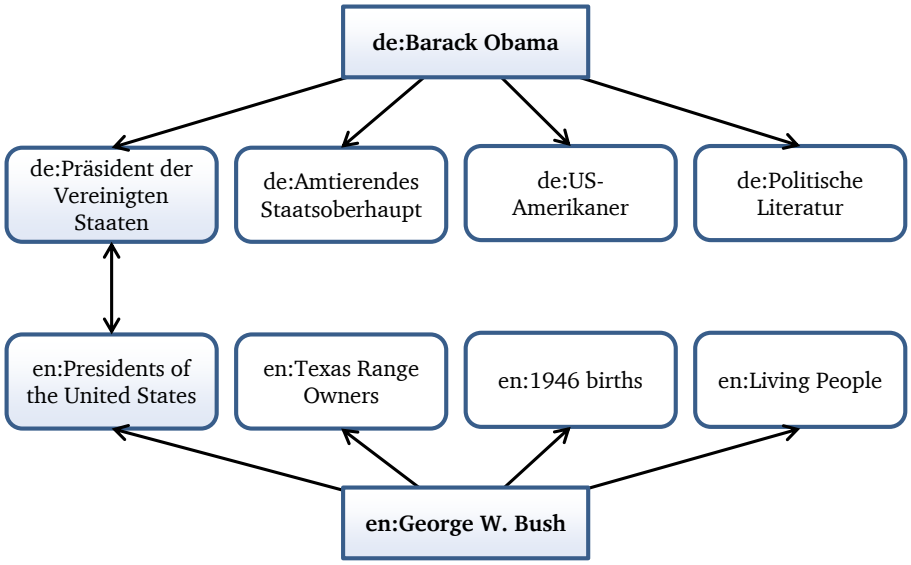


Figure 3.13: Example for Category CL Links. The example of two U.S.–American presidents shows that the category system of different languages is very different and the category overlap is scarce.

This specific example shows that, although there is a common category based on CL links, the applicability of this measure suffers from the different category structures which exist in different language versions of Wikipedia. For example, in the German Wikipedia, all persons are directly categorized as either *Mann*, *Frau* or *Intersexueller* (man, woman and intersex), whereas the English Wikipedia takes these categories as the roots for an intricate (and often inconsistent) category tree. So there often is no one–to–one mapping between categories. Further, in the German Wikipedia, 47% of all categories have an English counterpart, whereas only 7% of English categories have a correspondent German category. This unbalanced proportion has a negative impact on the applicability, as the probability of CL category overlap is low for most articles. However, some languages have a more compatible category structure that allows a good mapping. This is especially the case with languages that align their category structure with the English Wikipedia. As this is not the case with the German and English version, this approach is not applied in this thesis.

Usage of Meta CL Links

The Chain Link approach attempts to map an article written in the source language to a single article written in the target language which are both somehow related. For ESA, this means that the sparse semantic interpretation vector i_{esa} retains its density to a certain degree because a concept that has no correspondent concept in the target language is mapped to the best–matching concept. In situations where there is no matching concept, an only vaguely relevant or completely irrelevant mapping target

is chosen and noise is introduced. In the following, an approach is presented that does not attempt to match the best concept but tries to strengthen the already existing dimensions by adding Meta Cross–Language Links (MCLs). This approach infers a *meta–mapping* (in contrast to the direct mapping) from the article in the source language to a set of related articles in the target language. The weight of the article a_s in the source language l_s is split between all articles in the target language l_t that have corresponding articles in l_s linked to the article a_s. This is a fuzzy approach, but it has one transitive step less than the Chain Links approach. Therefore, the degree of abstraction is smaller with the MCLs than with Chain Links.

An example for such MCLs can be seen in figure 3.14. The English article *Clark Kent*[31] does not have a corresponding article in German. In order to map this concept into the German target language space, all the article's outgoing intrawiki links are checked for the existence of a German correspondent. The set of German articles that are connected by this way form a MCL and are considered to be related to the concept *Clark Kent*.

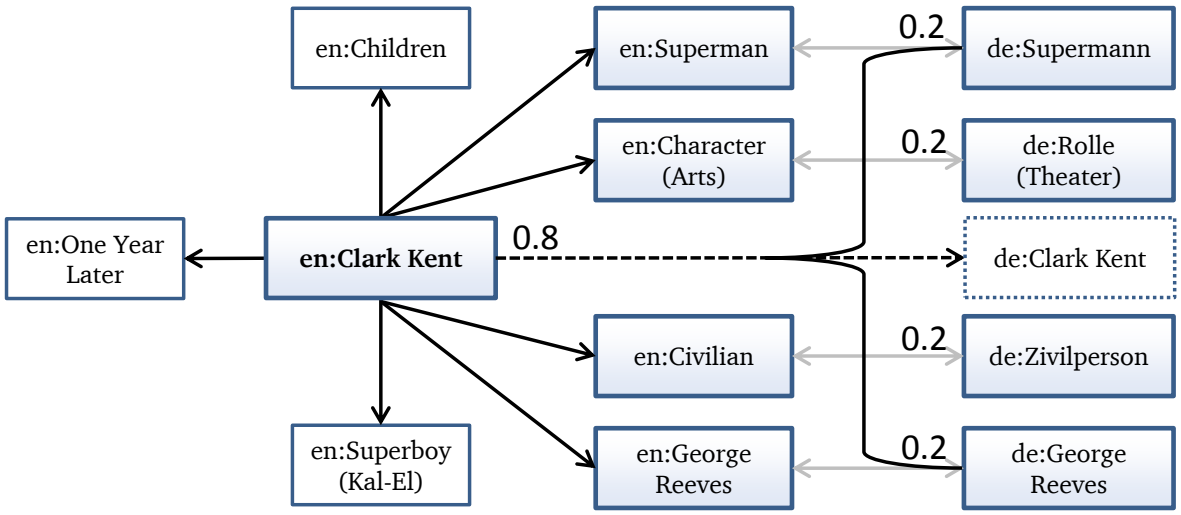


Figure 3.14: Example for a Meta Cross–Language Link. The missing German concept for *Clark Kent* is substituted by a normalized representation of all linked articles that have a respective translation in German. Here, the original value 0.8 of the concept *Clark Kent* is evenly distributed between all second degree CL links, thus each of the right concepts is strengthened with the value 0.2.

For cross–language ESA, the introduction of MCLs has implications on the dimensionality of the semantic interpreter M. All dimensions in the target language space that do not have an incoming CL link from the source language space do not have informative value in relation to the source language and therefore will not have any effect on the relatedness calculation. Thus, these dimensions are dispensable and can be removed completely. In practice, this means that for building the semantic interpreter M, articles that do not have a correspondent article in the other language can be ignored. If $L^* \subset L$ is the set of languages that is to be supported and c^{l_a} is the set of language independent concepts represented

[31] *Clark Kent* is the secret identity of super–hero *Superman*. `http://en.wikipedia.org/wiki/Clark_kent`, *retrieved 2011-02-25*

in each language $l_a \in L^*$'s semantic interpreter M^{l_a}, the size of the new reduced concept space that has to be investigated is given by

$$n_c = \left| \bigcap_{l_i \in L^*} c^{l_i} \right| \tag{3.12}$$

The dimension reduction can be performed by removing all the dimensions from our target concept space whose articles do not have an analogue in one of the other supported languages L^* of Wikipedia. Assuming the bidirectionality of CL links, this is simply the set of all articles that have an incoming CL link from articles for all other languages $\{l_x | l_x \in L^*\}$. In this thesis, however, only the use of two different languages is considered.

For the following evaluations, two different strategies are employed for mapping between the different languages:

The direct mapping just takes into account the existing CL links. The interpretation vector $i_{esa}^{l_s}$ in the source language l_s is mapped by removing all dimensions without a corresponding article in the target language l_t. The respective weights in $i_{esa}^{l_s}$ are directly transferred to $i_{esa}^{l_t}$ for all existing CL links.

The MCL mapping first transfers all weights from $i_{esa}^{l_s}$ to $i_{esa}^{l_t}$ where a CL link is existing. If no CL link exists, it is emulated using a MCL by deriving a set of articles $m = \{a_1^{l_s}..a_n^{l_s}\}$ linked by the source article and having a corresponding article in l_t. This set represents the *meta concept* of the source article a_s. The original weight is divided by the cardinality of m and added to the weight corresponding to $i_{esa}^{l_s}[a_n]$. Thus, the weight is equally distributed to all elements of the meta concept.

In the following, several evaluations are presented that compare cross–lingual ESA using direct mapping and MCL mapping.

3.5.3 Evaluations

For determining cross–lingual semantic relatedness, two evaluation types are presented in this section. The first part shows an evaluation of snippet comparison in an IR task. The second part focuses on relatedness of term pairs.

Evaluation of an Information Retrieval Task

Using the `Europarl300` subset of the `Europarl` corpus (cf. section 3.3.2) containing 300 parallel sentences in English and German, an IR task is performed. Due to computational constraints, only this small number of documents was chosen. In this evaluation, one document $d_q^{l_s}$ representing a sentence in the source language l_s is issued as a query and the parallel document $d_q^{l_t}$ in the target language l_t is expected as result, because the assumption is made that the semantic relatedness of this sentence pair is higher than for all other $\{d_n^{l_t} | n \neq q\}$.

There are three different experimental settings that are applied in this evaluation:

Direct Mapping is a mapping of the interpretation vector i_{esa} of the query document $d_q^{l_s}$ to the target language l_t using direct CL links.

MCL Mapping is a mapping of i_{esa} of the query document $d_q^{l_s}$ to the target language l_t using MCL.

Translation is a mapping of i_{esa} by translating the query document $d_q^{l_s}$ using Google Translator[32] to l_t and performing monolingual ESA there.

The `Europarl300` corpus was chosen, as to the best of the knowledge of the author, there is no available cross–lingual dataset that has similar properties like the German `Gr282` dataset (grouping related documents into *semantic groups*). Thus, the translation based approach is expected to outperform both mapping approaches significantly, as, ideally, the translation of a document is semantically identical to the parallel document in the other language.

The evaluations are performed in both directions, i.e. using a German source document as query to search for an English response and vice versa. This is due to the relevance of the direction of mapping: even if semantic relatedness is theoretically commutative, Wikipedia exposes different properties in different languages, e.g. the number of represented concepts or the quantity and generality of intrawiki links.

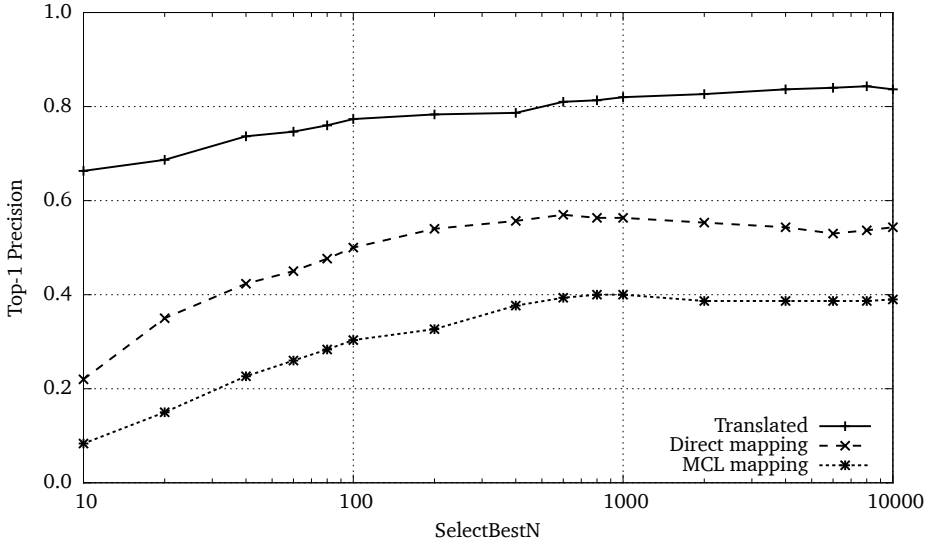
Figures 3.15a and 3.15b show the Top–1 precision of the implemented mapping strategies *direct mapping* and *MCL mapping*. Top–1 means that the IR task was only considered successful if the first returned result was the correct corresponding parallel sentence. For the Top–5 and Top–10 results which show to exhibit a similar trend, see figures A.2 and A.3 in appendix A.4.

As expected due to the quality of Google Translator, the translation based approach clearly outperforms both mapping approaches by far, yielding Top–1 precisions of 84% for the English target language space and respectively 80% for the German target language space, both at $n = 8000$. This is not surprising, as both interpretation vectors are computed in the same language and therefore are likely to have a similar concept weight distribution. The direct mapping shows to outperform the MCL mapping by large, yielding a maximum Top–1 precision at 57% for the English target language at $n = 600$ and 60% for German at $n = 10,000$, while the MCL mapping only achieves 40% for the English target language at $n = 800$ and respectively 26% for German at $n = 10,000$. Despite the seemingly large disparities of n between both languages, both mapping approaches reach a plateau of Top–1 precision between $400 < n < 1000$ for both languages.

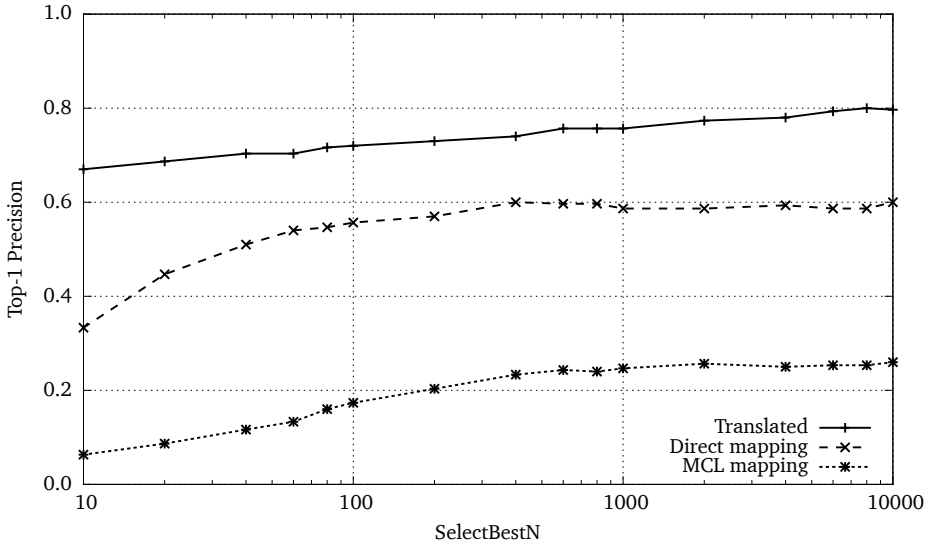
However, an interesting observation are the different results for direct mapping and MCL mapping when comparing both languages: while in the English target language space, the direct mapping has worse results than in German, the MCL mapping performs considerably better than its mapping to the German target language space.

The first effect can be explained by the different sizes of the respective Wikipedias: taking the assumption that a semantic interpretation vector of a query document is densely filled in its source language having s_s articles and it is mapped to a target language that has s_t articles, the following can be inferred: If $s_s < s_t$, the query's target language interpretation vector i_{esa_q} will be sparsely filled with at most s_s entries (if all articles in l_s have direct links from l_t). Thus, the mapped query interpretation vector will not have values for the concepts in l_t that do not have a direct link to l_s. Due to the applied cosine measure, these non–mappable values will introduce noise. However, if $s_s > s_t$, this noise is non–existent (again considering a full direct link mapping of the language with less articles) and therefore the similarities between mapped interpretation vectors and vectors in the target language are superior. For example, only 55% of German articles have a link to their English corresponding article, whereas only 18% of English articles link back. Therefore, a semantic interpretation vector that has been mapped from the German to the English language space can be filled by at most 18%, whereas the "native" vectors can be filled by up to 100%. Therefore, even when assuming that dense vectors never exist, the direct mapping

[32] `http://translate.google.de/`, *retrieved 2011-02-28*



(a) English Target Language Space l_{en}



(b) German Target Language Space l_{de}

Figure 3.15: Top–1 Precision for Information Retrieval task using the `Europarl300` dataset with disambiguation page filter and the three different mapping strategies. The x–axis represents the considered number of most relevant concepts of interpretation vector i_{esa}. The translation result is computed using monolingual ESA in the given target language space.

will degrade to some degree. As the MCL mapping relies on the direct mapping, as only existing direct links are strengthened, its results should degrade at the same rate. However, this is not the case.

There are two explanations for this difference: first, the MCL mapping is able to strengthen the relevant concepts in the English target language. This would mean that in the German Wikipedia, article links do not excessively link unrelated or too general articles, thus strengthening the "right" set of articles that describe relevant related concepts. This would hint on the superiority of the German article link topology. The second explanation is that the impact of the MCLs on the mapping is less drastic due to a lesser fraction of concepts that have to be mapped to the other language, thus diminishing the addition of noise in comparison to the direct mapping and therefore converging to the results of the direct mapping. A conclusive answer cannot be given in this thesis, but this will be a focus of future work.

The translation approach generally outperforms both mapping strategies. However, there are some exceptions, for example the translation of the sentence "There is no room for amendments" with the parallel sentence "Änderungen sind nicht möglich", which Google translates to "Changes are not possible". On calculating the semantic relatedness, there are other documents that are ranked before the correct parallel document. Both mapping approaches, however, are successful for this query document and return the parallel document as the first result. This example shows that for translations that do not match the original document precisely or contain only broad concepts, the cross–lingual approaches can be in advantage.

In conclusion, the MCL mapping is inferior to direct mapping for the comparison of CL documents. This is in part due to the addition of noise by spreading the article weights and in part due to the choice of the dataset `Europarl300`, as a one–to–one IR task benefits approaches that calculate semantic similarity. However, for a cross–lingual dataset that has similar properties like `Gr282` the results could improve considerably.

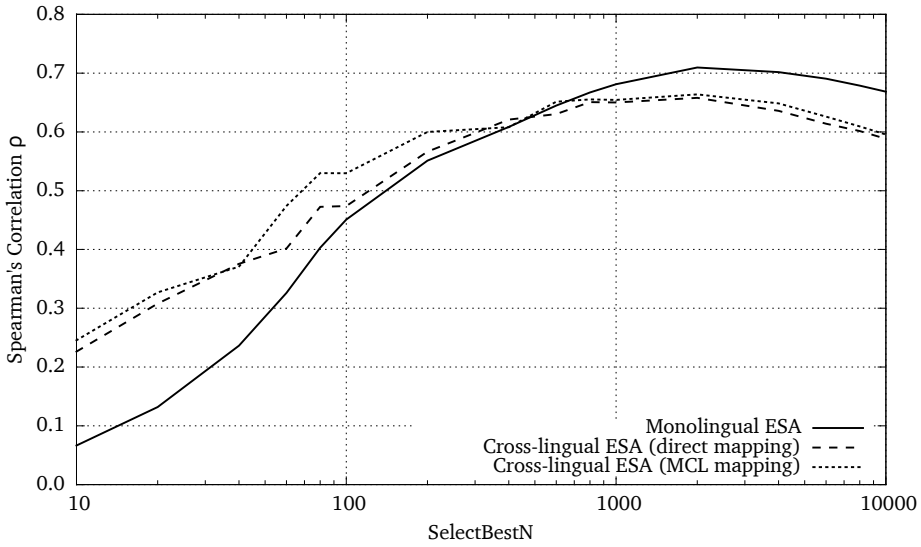
Evaluation of a Term Relatedness Task

Using the multilingual `Schm280` corpus (cf. section 3.3.2), the relatedness for (t_1, t_2) term–term pairings can be measured by calculating i_{esa} for each cross–language combination $(t_1^{l_1}, t_2^{l_2})$ in languages $l_1, l_2 \in \{l_{en}, l_{de}\}$ using direct mapping and MCL mapping. The respective target language space is the language of the first term. The resulting relatedness values are compared to the human ratings by calculating Spearman's rank correlation coefficient.

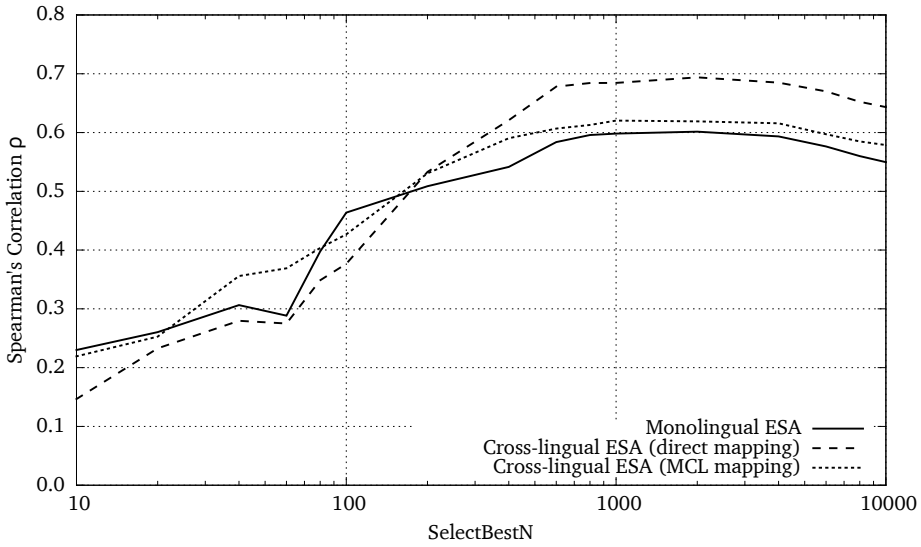
Figures 3.16a and 3.16b show the correlations between human raters' relatedness and relatedness determined by the cross–lingual direct mapping and MCL mapping in contrast to a monolingual relatedness calculation. In the monolingual setting, the Spearman's rank correlation coefficient ρ is maximal for *selectBestN* with $n \approx 2000$ for both languages (with $\rho_{en}(i_{esa}^{2000}) = 0.71$ and $\rho_{de}(i_{esa}^{2000}) = 0.60$). In cross–lingual settings, the experiment in the English language space l_{en} outperforms the experiment using the German language space by approximately 18% in terms of correlation.

When mapping the second term of each term pair from German to English by direct mapping, the correlation is significantly lower than the monolingual experiment (for $\rho_{en}(i_{esa}^{2000}) = 0.66$, $t_{\text{diff}} = 2.01$, $p < .05$). In the German language space, however, the direct mapping results in a significantly higher correlation (with $\rho_{de}(i_{esa}^{2000}) = 0.69$, $t_{\text{diff}} = 3.37$, $p < .01$) compared to the monolingual approach. This shows that the quality of ESA is highly dependent on the used language space and a cross–lingual mapping transfers qualitative properties to the target language space.

Applying MCL mapping does not show to improve the results. In the English language space, direct and MCL mapping perform similarly, and only with small n, the MCL mapping outperforms the direct mapping. This might be due to an accumulation of article links to relevant concepts (which usually are



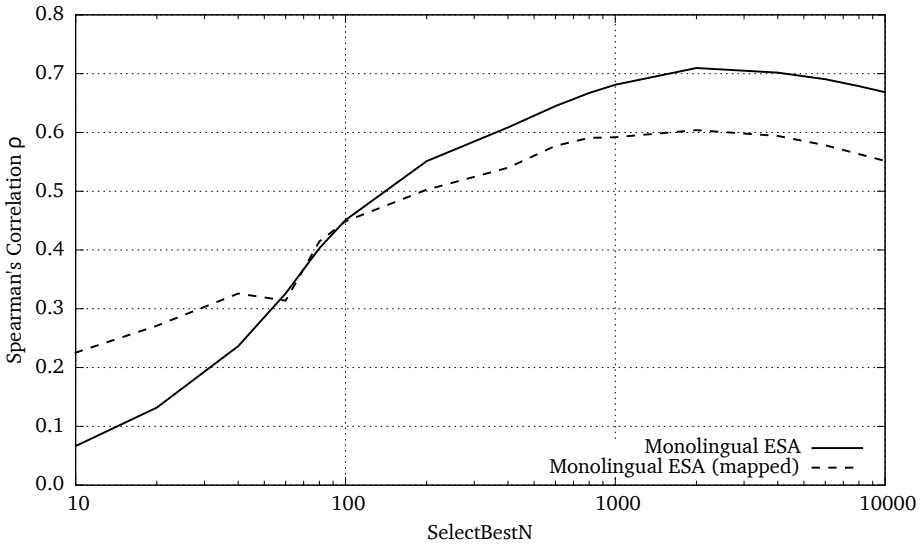
(a) English Target Language Space l_{en}



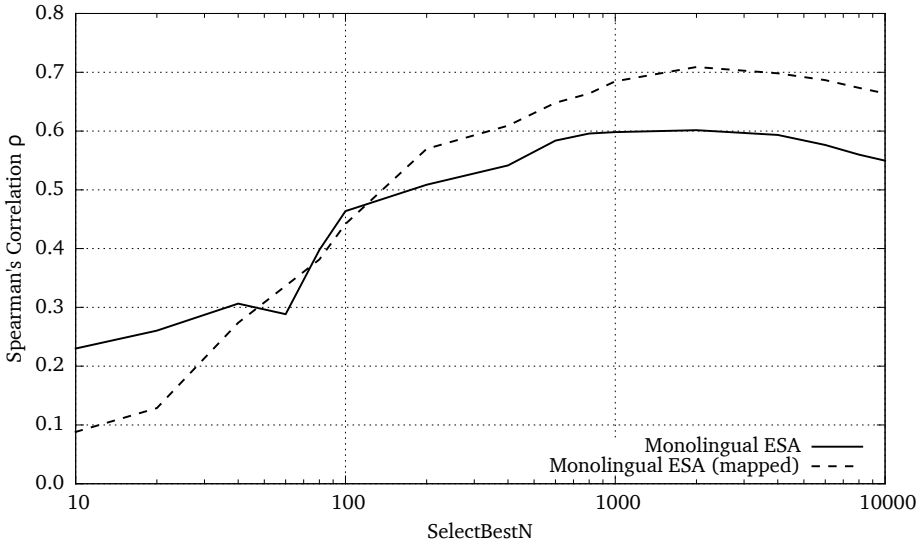
(b) German Target Language Space l_{de}

Figure 3.16: Correlation between human rated relatedness and relatedness determined by cross–lingual mapping strategies for `Schm280`

general and therefore have a statistically higher probability of being the target of article links), boosting the values for these relevant concepts due to spread. Manual analysis of the data shows in fact that the relevant concepts are benefiting from the MCLs' incoming article links. However, especially in the German language space, the correlation with MCL mapping is lower for high n.



(a) English Target Language Space l_{en}

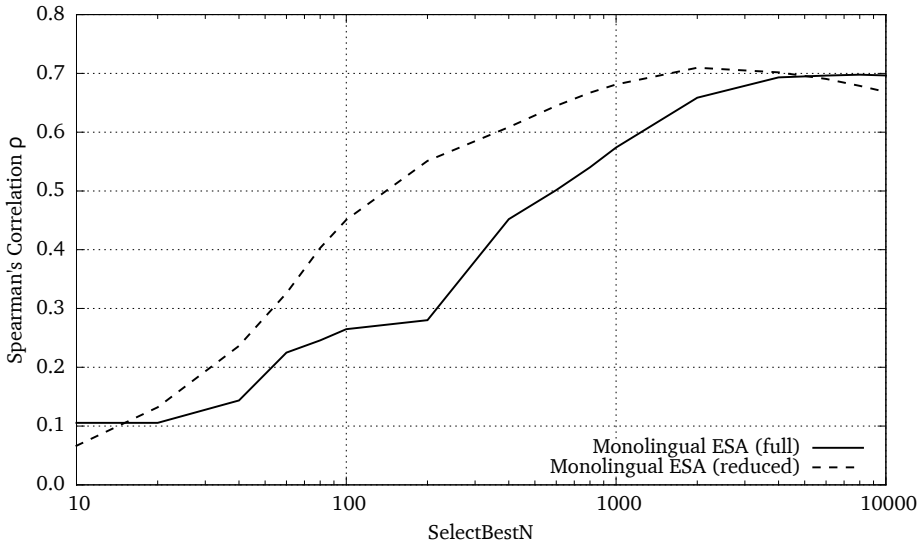


(b) German Target Language Space l_{de}

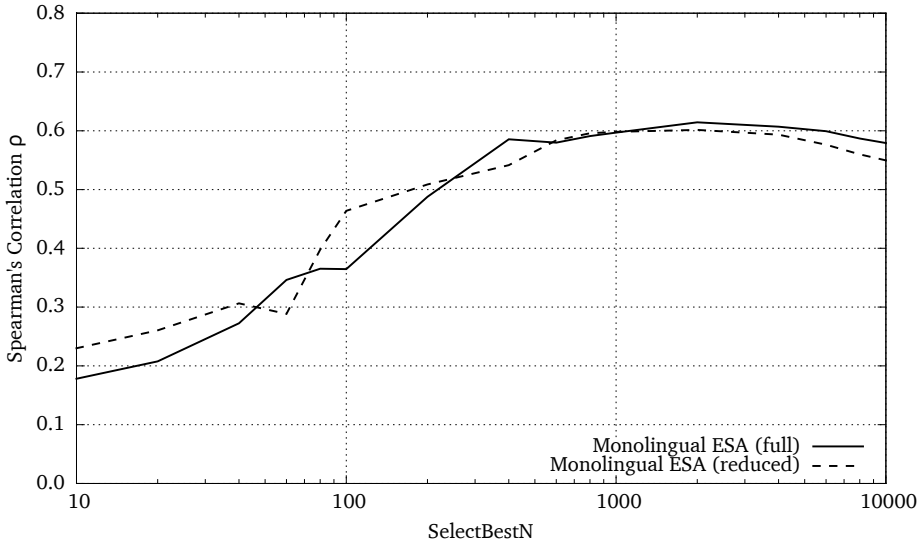
Figure 3.17: Correlation between human rated relatedness and relatedness determined by a reduced cross–language semantic interpreter. All disambiguation, category, portal, person and listing pages are filtered.

As shown in section 3.4, a concept space reduction is an applicable strategy to reduce the complexity of ESA while maintaining similar results. Figure 3.17 shows the results for both the English and German language space with the semantic interpreter M reduced by disambiguation, category, portal, person and listing pages. In comparison to figure 3.16, this reduction does not have a significant negative influence on correlation, in some cases it is even minimally improved. For example, in the German language space,

the maximum monolingual correlation increases from $\rho_{de}(i_{esa}^{2000}) = 0.60$ to $\rho_{de}(i_{esa}^{1000}) = 0.63$, although this difference is not significant.



(a) English Target Language Space l_{en}



(b) German Target Language Space l_{de}

Figure 3.18: Influence of changing the language space for both term pairs in a monolingual relatedness computation setting

The hypothesis that the language space is an important factor for the correlation is confirmed when inspecting the results of an experiment where *both* terms are transferred to the other language. In figures 3.18a and 3.18b, the results of such an experiment are shown using direct mapping. The main influence on the correlation is the language in which the interpretation vector is created, whereas the target language space has only a minor influence. Thus, when analysing the characteristics of figure 3.18 it can be inferred that the English language space is better applicable for the creation of ESA interpretation vectors than the German. This is supported by the hypothesis of Hassan et al. [86] that ESA's quality strongly correlates with the size of a language's Wikipedia in articles.

These findings can be utilized for a further article filter strategy (cf. section 3.4) that filters the semantic interpreter M's concept space by only taking into account articles that have a corresponding article in another given language space. This filter strategy is heavily dependent on the used languages but is not necessarily restricted to a multilingual relatedness setting. The reasoning behind this filter strategy is that for sufficiently large Wikipedias in different languages, CL links exist for the most important concepts. Concepts that do not have CL links are considered to be specific to language, location or culture of a certain language space and thus are not necessary for a generic semantic interpreter M in settings that do not need to take this specific information into account. Sorg and Cimiano [181] have already used this reduction strategy as a baseline for their CL–ESA approach implicitly, as their CL mapping ignores articles that do not have corresponding articles in the target language.

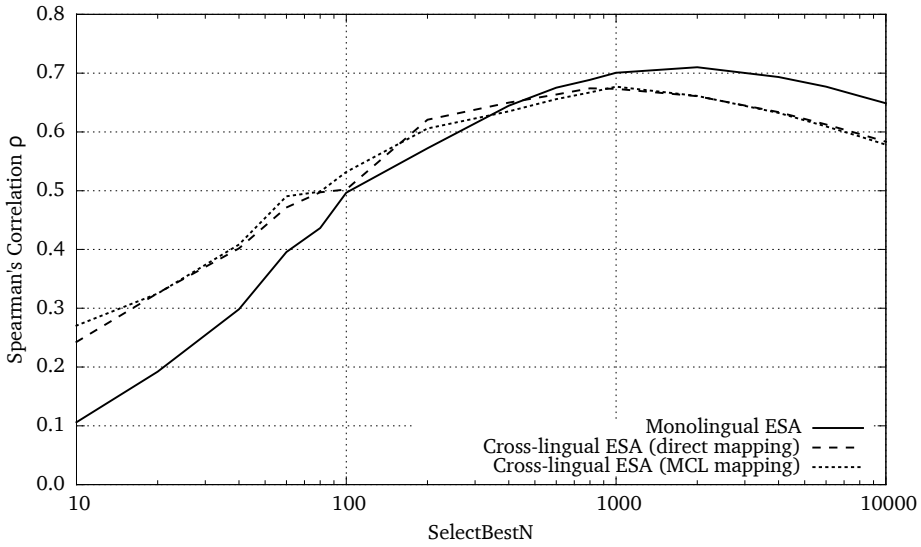
Figure 3.19 shows the correlation computed with the reduced semantic interpreter M by only taking into account articles that have a correspondent article in the other language in comparison to M using the full concept space. In the English language space, the choice of mapping methods does not have a major impact on the cross–lingual relatedness calculation, the correlation curves are similar for all n. However, with $n < 400$, the mapping from the German language increases the correlation in comparison with the monolingual ESA. This can be explained by the amount and weight of irrelevant concepts that are introduced with higher n in the English ESA by articles that are specific for the socio–cultural language area of the German Wikipedia. In the German language space, however, the direct mapping outperforms both the MCL mapping and the monolingual approaches for $n > 200$. On comparing the English and German monolingual ESA results, it should be noted that in general the English monolingual ESA seems to perform better than the German ESA. This shows that the introduction of a cross–language mapping can indeed increase the quality of a monolingual approach by a reduction of the concept space to articles existing in both languages. Further, it indicates that the quality of a semantic interpretation vector i_{esa} in a language that provides a good concept set can be partially transferred to other languages. As a consequence, it is to be examined whether the reduction of a semantic interpreter by removing concepts that have no correspondent article in another language is viable.

Figure 3.20 shows a comparison of different article filter strategies applied to monolingual ESA. The *NL filter* only accepts articles that have a corresponding article in a reference language. The maximum correlation is achieved with f1 : $\rho_{de}(i_{esa}^{1000}) = 0.63$ using the *CL* filter, which is not significantly different to the other filters (f2 : $\rho_{de}(i_{esa}^{2000}) = 0.62$ and f3 : $\rho_{de}(i_{esa}^{2000}) = 0.61$) but has a considerably reduced semantic interpreter M, having only 51.38% of the articles of the unfiltered semantic interpreter and 59.23% of the non–zero entries. Table 3.12 displays the results of the evaluation using other corpora.

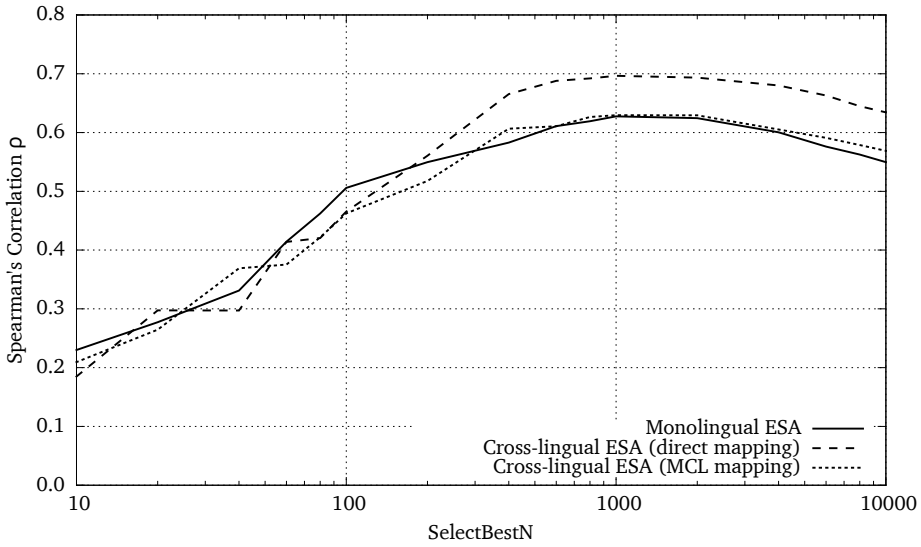
	RDWP984					Gur65	Gur350
Applied Filters	Covered	Wrong	Correct	Global Accuracy	Local Accuracy	Correlation ρ	Correlation ρ
None (all articles)	911	244	667	68.78%	73.22%	0.63	0.51
Language Filter English (NL_{en})	885	261	624	63.41%	70.51%	0.61	0.32

Table 3.12: Results of Cross–Language Reduction Strategy NL_{en} in a monolingual setting, filtering all German articles that do not have a directly corresponding English article. These results are derived from the original i_{esa} interpretation vectors.

The results using the RDWP984 dataset with the original interpretation vectors show that the coverage of terminology suffers from filtering articles that do not have a corresponding counterpart in the En-



(a) English Target Language Space l_{en}



(b) German Target Language Space l_{de}

Figure 3.19: Influence of reducing the concept space to cross–lingual concepts on monolingual correlation for English and German

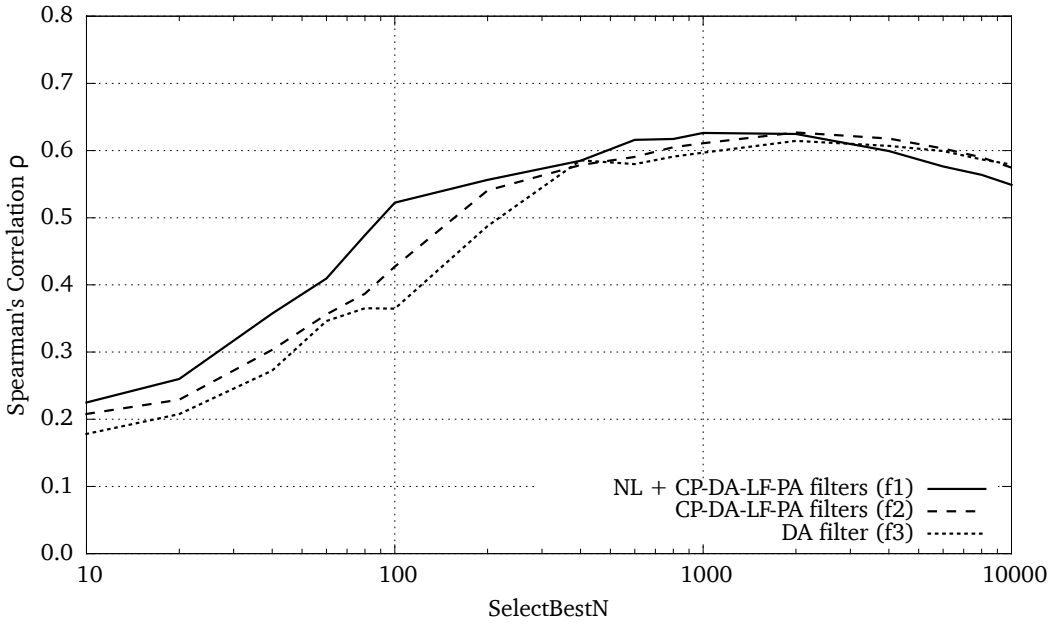


Figure 3.20: Influence of applying different article filtering strategies on correlation for monolingual relatedness evaluated in the German language space l_{de} using English as reference language space with `Schm280`.

glish reference language. This is due to the specific terminology in this dataset, containing very unusual and specific terms. The results of dataset `Gur65` are only marginally reduced, as the concepts that are contained there are general and likely to be contained in articles in both languages. `Gur350`, however, suffers significantly from the loss of articles. As this dataset covers very loosely related term pairs, it can be assumed that this strong reduction strongly affects the ability of the semantic interpreter to provide the necessary specific concepts that are used to infer such relationships. Therefore, for settings that require determining the relatedness of loosely related terms, the cross–language reduction strategy should not be used, in other settings it provides acceptable results while reducing the semantic interpreter M considerably.

3.5.4 Conclusions of Cross–Lingual ESA

In this section, the applicability of CL ESA has been shown for two selected scenarios and a novel mapping approach to enrich missing CL link information called *MCL mapping* has been presented. The focused application scenario involves short, cross–language documents in German and English that are commonly found in tagging systems like tags and snippets (cf. section 3.1.2), with the particular target ELWMS.KOM.

For snippets, a precision of up to 84% by using the direct mapping approach in a cross–lingual information retrieval task were shown. The MCL mapping did show to be inferior to direct mapping due to its introduction of noise in the target language. However, in some cases, the used translation engine could not enable the retrieval of the correct document, whereas the CL mapping approaches were able to do so. This shows the potential of cross–lingual mapping approaches relying on concepts instead of terms. For the evaluation, an evaluation corpus subset based on `Europarl` has been used which is arguably not perfectly adequate for evaluating approaches to semantic relatedness.

However, for term–term pairs (which are similar to tags in tagging systems), correlation between the relatedness determined by humans and by the presented approaches has shown to yield good results. On comparing the cross–lingual relatedness with the monolingual relatedness in German and English, it was shown that there are specific differences in the used language spaces of German and English. Especially using ESA in the English language space provides a more precise vector representation of the documents, resulting in less loss of information during the CL mapping.

The novel MCL mapping, however, in general did not improve the maximum correlation in comparison to direct mapping. This may be due to the fact that concepts that do not have corresponding concepts in the other language are either not highly relevant or that the dissemination of the source concept's weight to a set of concepts in the target language introduces a considerable amount of noise. A research question that has to be followed in future work is whether the MCL mapping may benefit from a *selective* distribution weighted based on the relatedness of the linked articles. Yet, for low n as parameter for the selection function *selectBestN*, MCL mapping showed to perform better than direct mapping by trend.

Eventually, experiments using the German and English Wikipedia show that a reduction of concept dimensions by filtering articles that have no corresponding article in the other language can be a viable article filter strategy. Specifically, the experiments showed that the correlation to human rankings does not decrease significantly while providing the benefits of smaller storage needs.

The transferability of the presented approaches depends on multiple factors:

- The size of the respective Wikipedia versions. As shown above, the cross–language approach works well for the two major Wikipedias in English and German. However, very small Wikipedias like the Gaelic Wikipedia[33] with approximately 12,500 articles are per se not an applicable corpus for ESA and therefore are equally inappropriate for the computation of cross–lingual semantic relatedness.
- The number of direct interlanguage links to the target language. As the foundation for all presented cross–lingual mapping approaches are the direct links, the respective Wikipedias have to include a considerable fraction of interlanguage links. For example, the Arabic Wikipedia contains approximately 140,000 content articles with 52.55% of the articles having a direct link to the German and 78.73% having a direct link to the English Wikipedia. This suggests that most Wikipedia languages are well interlinked to the major Wikipedia languages. Users of Wikipedia support the interlinking of different languages by providing the Interwiki–Bot[34] which regularly crawls the different Wikipedias to maintain the CL link structure by inferring over existing direct links. Thus, it can be assumed that the CL link structure is quite complete.
- The generality of concepts that have a corresponding article in the target language. A requirement of the CL mapping is the availability of interlanguage links between relevant concepts for the scenario. If this requirement is not met, the quality of CL ESA will suffer. Fortunately, at least the major Wikipedias provide CL links for the most important concepts.

Thus, it always depends on the specific usage scenario, whether the presented approaches can be transferred to other languages.

3.6 Extended Explicit Semantic Analysis

As described in section 3.2.3, the original ESA makes only use of the article information that Wikipedia contains, i.e. the term → article allocation. However, Wikipedia provides a wealth of additional semantic

[33] `http://ga.wikipedia.org/wiki/Pr%C3%ADomhleathanach`, *retrieved 2011-04-07*
[34] `http://en.wikipedia.org/wiki/User_talk:Yurik/Interwiki_Bot_FAQ`, *retrieved 2011-04-07*

information, respectively the links between articles and the categorization structure of articles. The original ESA approach neglects this information completely.

Thus, in the following, eXtended Explicit Semantic Analysis (XESA) is introduced as an approach that semantically enriches the semantic interpretation vectors i_{esa} obtained from ESA. In detail, three different approaches to extending ESA are presented: one utilizing the article link graph of Wikipedia, one using the category structure and one approach that combines those two. The basic idea is to enrich the interpretation vector i_{esa} with additional semantic information that can be extracted from the Wikipedia corpus. Due to this additional intrinsic semantic information, a better quality of the approach is expected.

3.6.1 Utilization of the Article Graph

On average, each German Wikipedia article links to 31 other articles. These article links can be interpreted as semantic relationships to other concepts. For example, the German article for *General Relativity* links to the other articles *Space*, *Time* and *Gravitation*. Thus, there is an obvious generic relatedness to the concepts expressed by these article links. The utilization of the article graph is an extension to ESA that aims at benefiting from the *associative* semantic information contained in Wikipedia article links.

The overall article linkage graph of Wikipedia can be represented as the adjacency matrix $A_{Articlegraph}$ of dimensions $art \times art$, where art is the number of articles contained in the semantic interpreter M. If an article a_i links to a_j, the respective cell in the matrix is set to one, otherwise it is set to zero (cf. equation 3.13), resulting in a highly sparse matrix that is typically filled by less than 0.01‰.

$$A_{Articlegraph_{i,j}} = \begin{cases} 1.0 & \text{if } a_i \text{ contains a link to } a_j \\ 0.0 & \text{otherwise} \end{cases} \tag{3.13}$$

Only directly neighboured articles are taken into account. Another approach could be to include weights that decrease with the linkage distance of articles on indirect links, e.g. if a_i links to a_j and a_j links to a_k, that a value greater than 0 (but less than 1) is inserted into $A_{Articlegraph_{i,k}}$. Yet, a closer examination reveals that the semantic relatedness between articles linked by second degree is already very low, thus it would only raise computation overhead without contributing to the result. For example, the article *General Relativity* links to *Space*, which again links to the articles *Knowledge* and *Measurement*. In the latter two articles, however, there is no information that adds to a semantic description of the concept of General Relativity. Therefore, this weighted measure based on the article linking distance is not applied.

As articles usually do not contain references to themselves, the adjacency matrix has to be added to the identity matrix I_{art} so that the diagonals are not zero. Otherwise, there is the possibility that already computed information is lost. Further, a weight factor w is introduced that determines how strong the influence of the article graph is on the original i_{esa}. Multiplying ESA's semantic interpreter M with the resulting matrix (equation 3.14) reinforces relevant semantic information and introduces new information based on the article linkage. The result is the new semantic interpretation vector $i_{xesa^{ag1}}$.

$$i_{xesa^{ag1}} = i_{esa} \cdot (w * A_{Articlegraph} + I_{art}) \qquad w \in [0..1] \tag{3.14}$$

Performance–wise, this article graph extension poses the challenge that the complete interpretation vector has to be multiplied with a large matrix again. As i_{esa} usually contains only few similarity values

that are significant and lots of values that are really small, the function *selectBestN* is applied for boosting efficiency of calculation (cf. section 3.4.6) and i_{esa} is truncated after the first best n similarity values. This has the effect that the second matrix multiplication is more efficient to be calculated because i_{esa} is only sparsely filled with values > 0. Thus, a second version of i_{xesa} is defined that reduces the overall calculation complexity by only taking the n highest similarity values into account (3.15).

$$i_{xesa^{ag2}} = i_{esa} + selectBestN(i_{esa}, n) \cdot (w * A_{Articlegraph} + I_{art}) \qquad w \in [0..1], n \in \mathbb{N} \tag{3.15}$$

A challenge, though, is finding an appropriate n that speeds up calculation without deteriorating the quality of the result too considerably. This issue is dealt with empirically later in this section.

3.6.2 Utilization of Category Information

The category structure of the German Wikipedia contains approximately 100,000 categories with about 920,000 articles categorized (i.e. approximately 87% of all articles). Besides administrative categories and categories that group different articles by properties of the underlying concepts (e.g. *list of German authors by birth year*), there are categories that represent groupings by semantic properties and express (among others) *is–a* relations. Especially categories of this relationship are interesting for the enrichment of semantic interpreters, as they can strengthen the weight of concepts that are in the same category. Thus, the utilization of the category structure serves to tap the *hierarchical* semantic information contained in Wikipedia.

In order to achieve this, information that encodes category affiliation is appended to the interpretation vector i_{esa}, similar to [78]. Therefore, the category matrix C is created with the dimensions $n \times m$, where n is the number of articles and m the number of categories (see equation 3.16).

$$C_{i,j} = \begin{cases} 1.0 & \text{if article } a_i \text{ is a direct child of category } c_j \\ 0.0 & \text{otherwise} \end{cases} \tag{3.16}$$

This matrix is applied to the interpretation vector with the result being the vector $c_{cat} = i_{esa} \cdot C$ that encodes information about articles and categories. Finally, the resulting vector c_{cat} is appended to the semantic interpretation vector and the XESA category vector $i_{xesa^{cat}} = (i_{esa}, c_{cat})$ is obtained (the appending operator "," denotes that the second vector is suffixed to i_{esa}).

$$i_{xesa^{cat}} = (i_{esa}, selectBestN(i_{esa}, n) \cdot C) \qquad n \in \mathbb{N} \tag{3.17}$$

This operation increases the dimension of the vector i_{esa} by the number of category vector dimensions. Analogue to the approach using the article graph, this calculation is inefficient if all non–zero values are kept; thus, *selectBestN* is applied to i_{esa} again, resulting with equation 3.17.

3.6.3 Combination of Article Graph and Category Extensions

Finally, the article link and category extensions to ESA can be applied in combination. This extension aims at combining the associative information introduced by the article graph and the hierarchical information derived from the category structure. The definition of this extension is rather straight–forward, instead of i_{esa} the result of the article graph extension $i_{xesa^{ag1}}$ is used (equation 3.18).

$$i_{xesa^{combination}} = (i_{xesa^{ag1}}, selectBestN(i_{xesa^{ag1}}, n) \cdot C) \qquad n \in \mathbb{N} \tag{3.18}$$

This approach is computationally expensive, as it involves two matrix multiplications for each interpretation vector i_{esa}. Therefore, *selectBestN* is applied in the multiplication with the category matrix C. However, a danger of this double enrichment and reduction is the dilution of the original interpretation vector. Now that both associative and hierarchical information are utilized, this may add considerable noise in addition. For that reason, this combination is not expected to perform better than the article graph extension or the category extension on their own.

3.6.4 XESA Evaluations

In all of the following evaluations, the German Wikipedia dump from June 04, 2009 is used to build the semantic interpreter M. Two different settings are used for the experiments using XESA, one determining the semantic relatedness of term pairs using the `Gur65` and `Gur350` datasets and one for documents using the `Gr282` corpus (cf. section 3.3.2 for dataset descriptions).

Evaluation of XESA for the Relatedness of Documents

The `Gr282` dataset is used as a corpus for the following experiments. For a semantic interpreter M using the original ESA parametrization (cf. section 3.4), the BEP is at 0.575, the MAP is 0.595 with standard deviation 0.252.

Figure 3.21 shows the precision–recall diagram of ESA applied on `Gr282`. As precision and recall have an inverse relationship, the curve is always declining from left to right. The target of any approach to enhance ESA is to ensure that both precision and recall are increased at the same time. Thus, the slope of the curve indicates the quality of ESA. The more the curve bulges towards the right upper corner, the better the extension to ESA is.

Empirical Evaluation of selectBestN and Article Graph Weights

As described in section 3.6, the function *selectBestN* is introduced that discards all i_{esa} values but the n best values for better calculation performance. As all experiments using the different variants of XESA showed similar results in the impact of the choice of n, here only the XESA variant using the article graph ($i_{xesa^{ag1}}$) with three different values ($n \in \{10, 25, 100\}$) is shown in figure 3.22.

The results in figure 3.22 show that the article graph extension performs best with $n = 25$. This means that for the examined parametrizations, the best 10 concepts are too few to describe the documents' contents. With an increasing n, the results become better, but too many concepts introduce noise to the results. As this is consistent with the results obtained using the other extensions as well, in the following, only results are presented that are computed with $n = 25$.

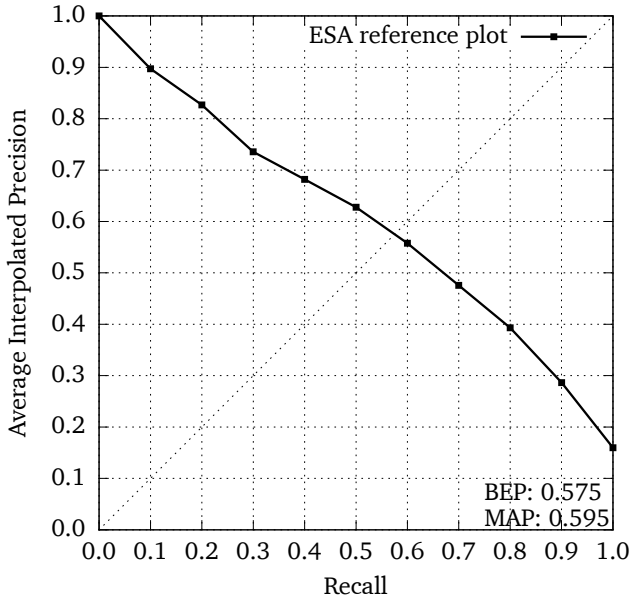


Figure 3.21: The precision–recall diagram for `Gr282` dataset using basic ESA with the Break Even Point where $f(r) = r$

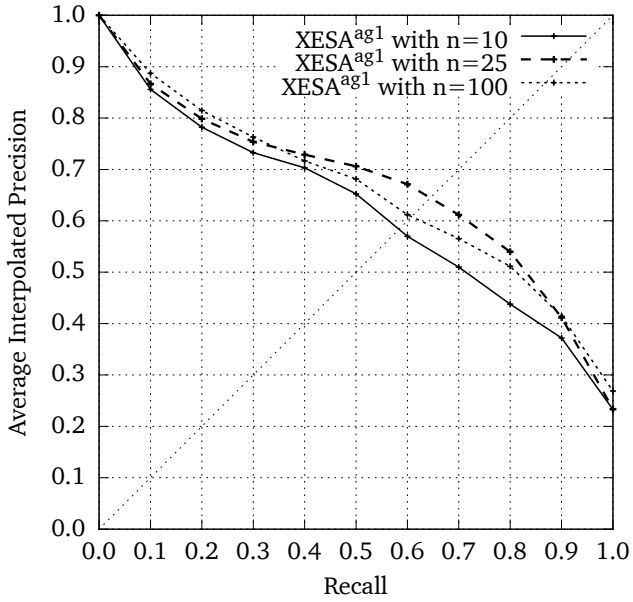


Figure 3.22: Impact of the semantic interpretation vector reduction strategy *selectBestN* with the article graph extension using $n \in \{10, 25, 100\}$

In contrast to the *selectBestN* results presented in section 3.5, n in XESA has to be considerably lower in order to achieve the best results. This is because these extensions to ESA tend to strengthen relevant concepts, making them more probable to be ranked highly.

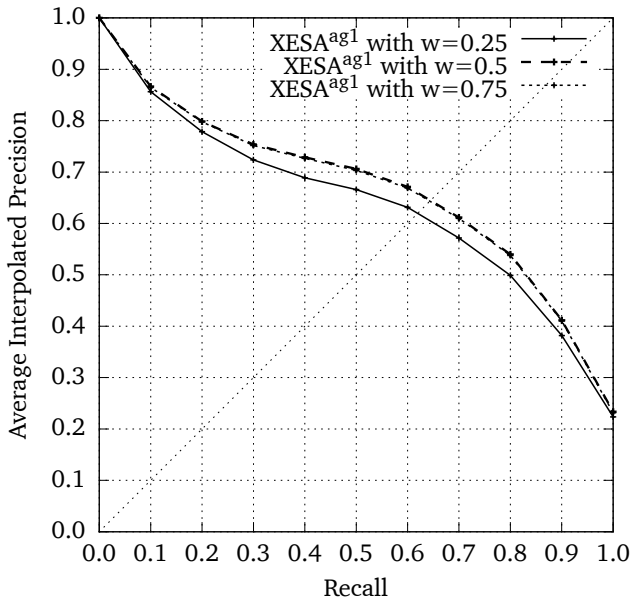


Figure 3.23: Impact of using different weights $w \in \{0.25, 0.5, 0.75\}$ for the article graph extension

Further, the article graph weight w used with all XESA article graph extensions was tested with $w \in \{0.25, 0.5, 0.75\}$ (see figure 3.23). In the experiments, the precision–recall curves for the weights $w \in \{0.5, 0.75\}$ are nearly identical, whereas $w = 0.25$ is already too small for having a considerable impact when multiplying the article graph matrix $A_{Articlegraph}$. Therefore, $w = 0.5$ is used in all following results.

Comparison of ESA and XESA

In this section, we compare results of the different XESA variants presented in section 3.6.

The precision–recall diagrams of all XESA variants using the *selectBestN*–parameter $n = 25$ and the link article graph weight $w = 0.5$ are displayed in figure 3.24. This plot shows that both article link graph extensions perform best, surpassing ESA results by 7%, whereas the category extension still outperforms ESA by 5.4% but cannot measure up to the article graph variants. Both variants combined, however, are not able to even achieve the performance of the basic ESA approach. Detailed results are additionally displayed in table 3.13.

Approach	Break Even Point	Mean Average Precision	Standard Deviation
ESA	0.575	0.595	**0.252**
XESA $i_{xesa^{ag1}}$	0.646	0.654	0.286
XESA $i_{xesa^{ag2}}$	**0.646**	**0.658**	0.284
XESA $i_{xesa^{cat}}$	0.629	0.647	0.274
XESA $i_{xesa^{combined}}$	0.539	0.515	0.301

Table 3.13: Summary of XESA's results (best are marked bold)

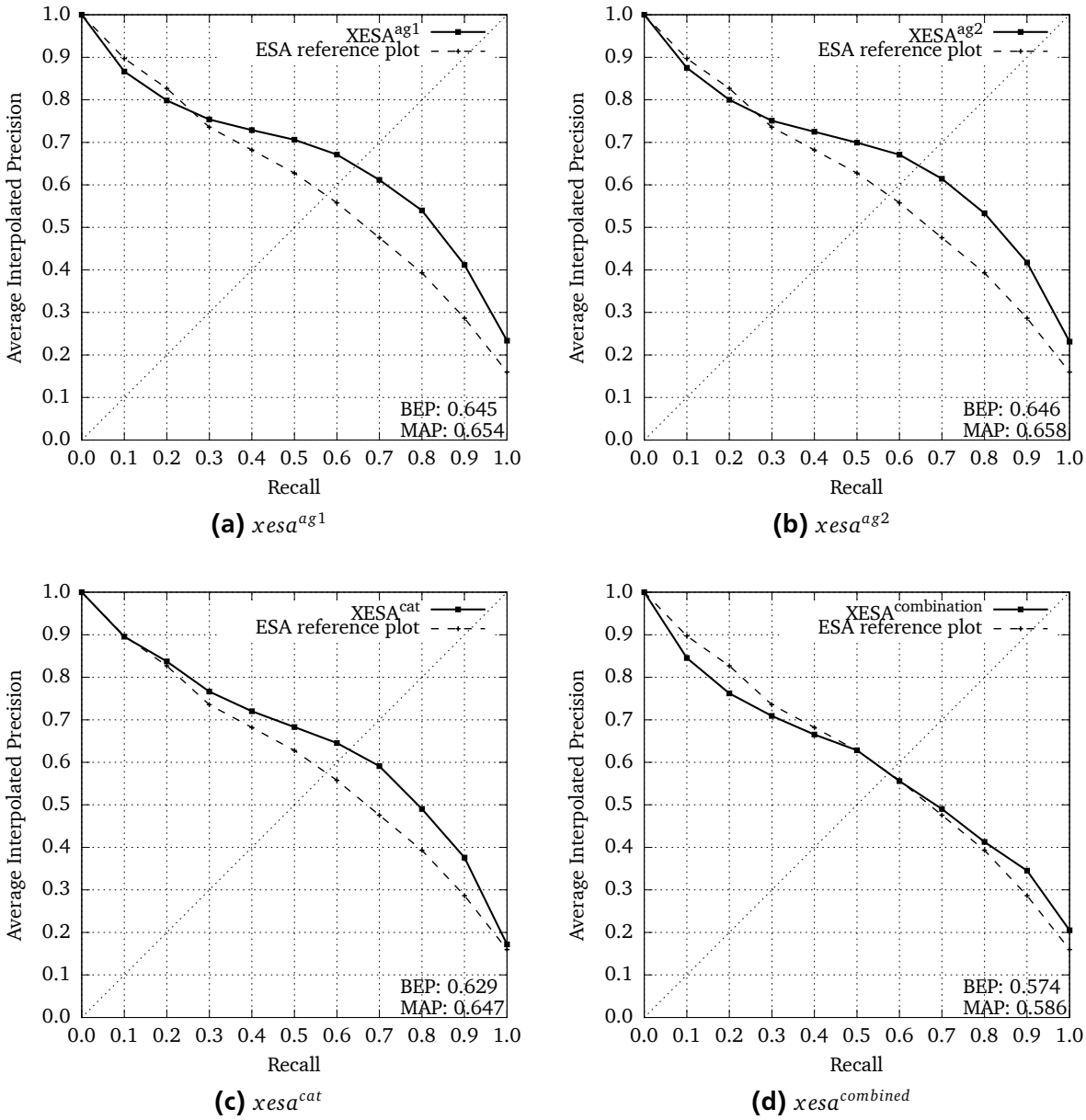


Figure 3.24: Precision–Recall plots of all XESA variants

These results show that the semantic information that can be derived from the Wikipedia article graph and the categories is beneficial for computing the semantic relatedness between documents. The article graph variants of XESA perform best because they represent a specific, associative relatedness between concepts. By linking articles, the human creators of the article want to express closeness of the underlying concepts. While being linked, some context of this relation can also be found in the linking article as well. For example, the article *General Relativity* links to the article *Space* and shares terminology with that article. Thus, by adding information about the relation, semantic information already known is strengthened by this connection. Categories, however, provide an organizational, top–down view on the concepts. While they provide semantic information about the grouping of articles, they are already abstracted from the specific concept itself. Therefore, the results of XESA's category variant improve ESA but still cannot measure up to the article graph variants.

Further, the results of the XESA combination variant are worse than ESA's results, probably because a multiplicatory effect occurs. By multiplying the interpretation vectors of different semantic dimensions in that approach, a semantic diversification occurs, i.e. the interpretation vector $i_{xesa^{combination}}$ is enriched by semantic information based on heterogeneous sources (article graph and categories). Thus, noise is added and the specificity of the semantic dimensions is decreased considerably.

As expected, the 14 semantic groups of the corpus proved to perform differently based on their abstraction. For example, snippets containing fact knowledge in a narrow topic are more easily related than broad topics, because certain terms are common in that group. XESA showed to outperform ESA in recognizing the semantic relatedness between documents in the groups that use different terminology.

Evaluation of XESA for the Relatedness of Term Pairings

Using the datasets `Gur65` and `Gur350`, XESA was compared to ESA with regard to the correlation of semantic relatedness of term pairs and human judgements.

On `Gur65`, the original ESA results in the Spearman's rank correlation coefficient $\rho = 0.678$. This value is the baseline for comparing the XESA extensions.

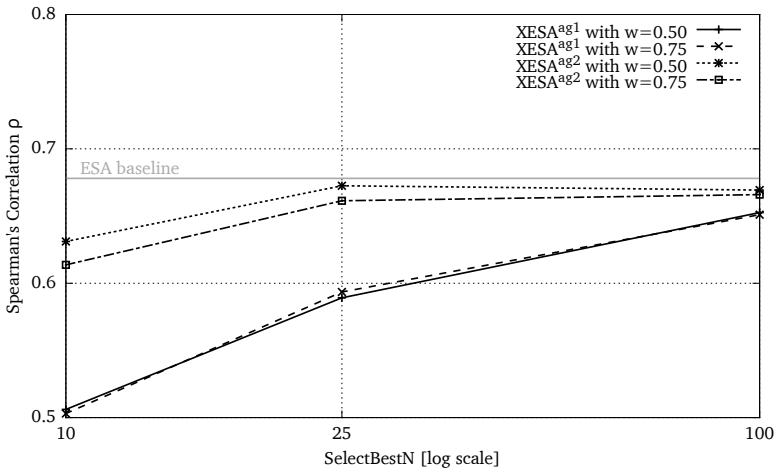


Figure 3.25: XESA article graph extensions XESA^{ag1} and XESA^{ag2} parametrized with different $n \in \{10, 25, 100\}$ and article weights $w \in \{0.5, 0.75\}$ on `Gur65` dataset

First, the parametrization of *selectBestN*'s n and the article graph weight w are validated using the article graph variants of XESA ($i_{xesa^{ag1}}$ and $i_{xesa^{ag2}}$) exemplarily. Figure 3.25 shows that XESA^{ag2} performs continuously better than XESA^{ag1}, but neither of both variants is able to best the original ESA approach. Analogue to the experiments using `Gr282` in order to determine the article graph weight w, different settings for w do not change the results considerably. Again, a good value for the weight is $w = 0.5$, which is subsequently used in the following evaluations. For n there seems to be a difference between the two corpus types. Where the document comparison performs best with $n = 25$, for a term–term relatedness computation a higher n seems to perform better by trend. However, further experiments show that the choice of n does have a varying impact on the results.

Figure 3.25 shows the correlations the XESA extensions achieved and the ESA baseline. Here, the varying influence of the choice of *selectBestN*'s n can be seen. Whereas the pure article graph extensions benefit from a higher n, XESA^{cat} and XESA^{combined} are clearly impaired by inclusion of more high–ranked concepts. This is due to the enrichment of the semantic interpretation vector i_{esa}. For the semantic

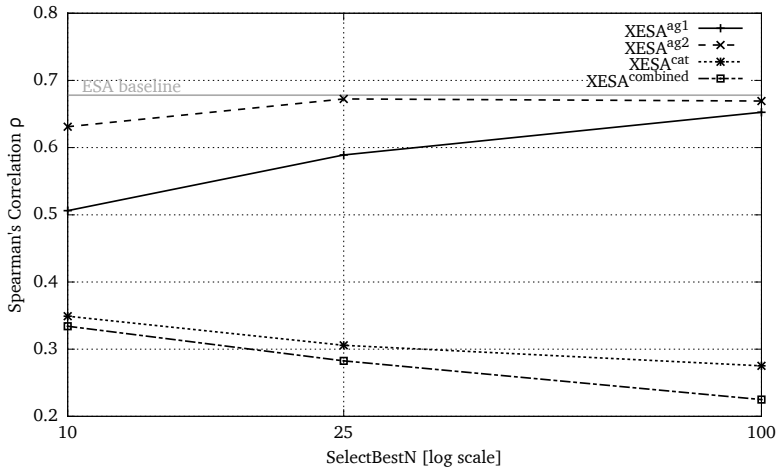


Figure 3.26: Performance on `Gur65` dataset of all XESA extensions parametrized with different $n \in \{10, 25, 100\}$

relatedness of terms, the article graph extensions XESA^{ag1} and XESA^{ag2} seem to strengthen the relevant concepts in i_{esa}, whereas a further enrichment via XESA^{cat} and XESA^{combined} introduces too much noise, reducing the correlations between the approaches and the human judgements considerably to a level where the correlations can be interpreted as weak.

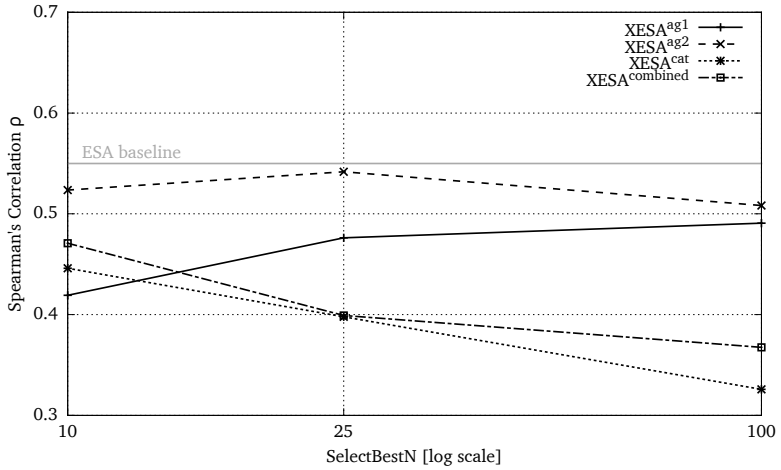


Figure 3.27: Performance on `Gur350` dataset of all XESA extensions parametrized with different $n \in \{10, 25, 100\}$

A similar effect can be seen for the dataset `Gur350`, where the baseline of ESA is $\rho = 0.55$. Figure 3.27 shows the performance of all different XESA extensions, with XESA^{ag2} being the approach that mirrors the human judgements best without achieving the accuracy of the original ESA approach.

The collection of best XESA approaches are shown in table 3.14. Again, no XESA extension can achieve the results of ESA, but XESA^{ag2} comes close. This is different from applying XESA on semantic relatedness of short documents, where especially enriching i_{esa} with associative information based on the article graph shows to increase the performance. In this case, XESA is able to exploit the additional semantic information contained in Wikipedia. This may be due to the context that is given by additional terms in documents. For single terms, however, the ESA approach seems to provide already all information that is needed to provide an applicable semantic relatedness computation. Further enriching i_{esa} does introduce

	Gur65 dataset		Gur350 dataset	
Approach	Best Parametrization	Correlation ρ	Best Parametrization	Correlation ρ
ESA	-	0.678	-	0.550
XESA^{ag1}	$n = 100, w = 0.5$	0.653	$n = 100, w = 0.5$	0.491
XESA^{ag2}	$n = 25, w = 0.5$	0.673	$n = 25, w = 0.5$	0.542
XESA^{cat}	$n = 10, w = 0.5$	0.349	$n = 10, w = 0.5$	0.446
XESA^{combined}	$n = 10, w = 0.5$	0.334	$n = 10, w = 0.5$	0.471

Table 3.14: All results of comparing the correlation for ESA and XESA using the datasets Gur65 and Gur350

noise and, therefore, degrades the results. Thus, this scenario with single term lacking additional context does not benefit from XESA, contrary to the semantic relatedness computation of documents.

3.6.5 Conclusions of XESA

In this section, several ESA extensions have been presented that incorporate additional associative and hierarchical semantic information contained in Wikipedia, namely the article graph and the category structure of Wikipedia. The impact of these extensions (named XESA) were shown for two different evaluation settings. The results indicate that XESA, especially the article graph extension, is beneficial for settings that involve the comparison of multi–term documents, whereas for single–term documents, XESA could not match the original ESA approach.

3.7 Conclusions

In this chapter, a scenario of how recommendations for ELWMS.KOM can benefit from inferring relatedness of documents was presented. The measure of relatedness is more suited to such a task than similarity, as learners do not only need to be recommended similar LRs containing information they might already know but also related LRs that provide new insights or a novel perspective on the learning matter they actually work on. An analysis of user data from ELWMS.KOM showed that the challenge of employing semantic relatedness in ELWMS.KOM is that tags and LRs are usually short, do not expose much exploitable context and are commonly composed in different languages. Related work was analysed on the basis of these properties and ESA was identified as a basic approach that conforms to the requirements. The novel semantic corpus Gr282 was presented that adequately represents the nature of LRs in ELWMS.KOM. Several strategies were proposed and examined that aim at reducing the computational complexity of ESA while retaining its precision. Further, it was shown that some of the reduction strategies are able to reduce noise and therefore can boost ESA's results. Additionally, the applicability of ESA for cross–language semantic relatedness was shown, introducing a new strategy to overcome missing CL links in Wikipedia. Eventually, three novel approaches of semantically enriching the interpretation vectors obtained by ESA based on Wikipedia article links and categories were presented. These extensions, subsumed under the name XESA, were evaluated and it was shown that the extension based on the article link graph outperforms ESA by 7% on the novel corpus of snippets Gr282. XESA was examined for semantic relatedness of documents and single terms, showing the approach's limitations concerning noise in term–pairing comparisons. Yet, it can be inferred that ESA, albeit already a stable and qualitatively good approach, can be enhanced in a document–based retrieval scenario by using further semantic information contained in Wikipedia.

For ELWMS.KOM this means that computing semantic relatedness as a basis for providing content–based recommendations is feasible. It has been shown that the presented approaches work for single terms and short snippets in mono– and cross–lingual settings. An advantage of XESA in comparison with other approaches (e.g. LSA) is that it is able to show the learner the concepts that are computed to be relevant for a LR, introducing transparency and effectiveness (cf. [188]) to the recommendation process. As these concepts correspond to Wikipedia articles, the learner can be referred to them for a conceptual clarification, essentially utilizing Wikipedia as an additional source for LRs.

In future work, the focus will be on the recommendation engine that provides semantically related content. An open question is, whether and how learners benefit from the offering of unknown, but related, snippets. In the sense of the *serendipity effect*[35] [75], an interesting research question will be whether learners profit more from strongly or weakly related LRs. This requires further evaluations in an open self–directed learning setting using ELWMS.KOM.

Further, a challenge will be the question whether Wikipedia lemmata — the titles of the articles — may serve as human–readable topical hints and even as recommended tags for learners.

Eventually, the XESA approach introduced the inclusion of further semantic information from Wikipedia and opens a lot of successive research questions. For example, taking into account the *relevance of links* between articles could further improve the article graph extension as well as the MCL mapping. For example, the article *General Relativity* links to *Baltimore*, which is less relevant than the link to *Spacetime*. A weighting scheme that aims at reflecting this different degree of relatedness could significantly enhance the presented XESA and cross–lingual approaches. This is an interesting research topic to be covered in future work.

[35] The serendipity effect is based on the observation that often, relevant information is found by aimless browsing and chance, and not by a targeted search.

4 Granularity of Web Resources

The granularity of a LO is an important focus of TEL research, as LOs are expensive and laborious to create and strategies to efficiently re–use existing LOs is essential [130]. Especially the paradigms of *authoring by aggregation* and *re–purposing* of LOs [92, 116, 161, 205] are fields of research that build on the availability of multi–granular LOs. The granularity aspect is mainly targeted at supporting the authoring process in the lifecycle of LOs.

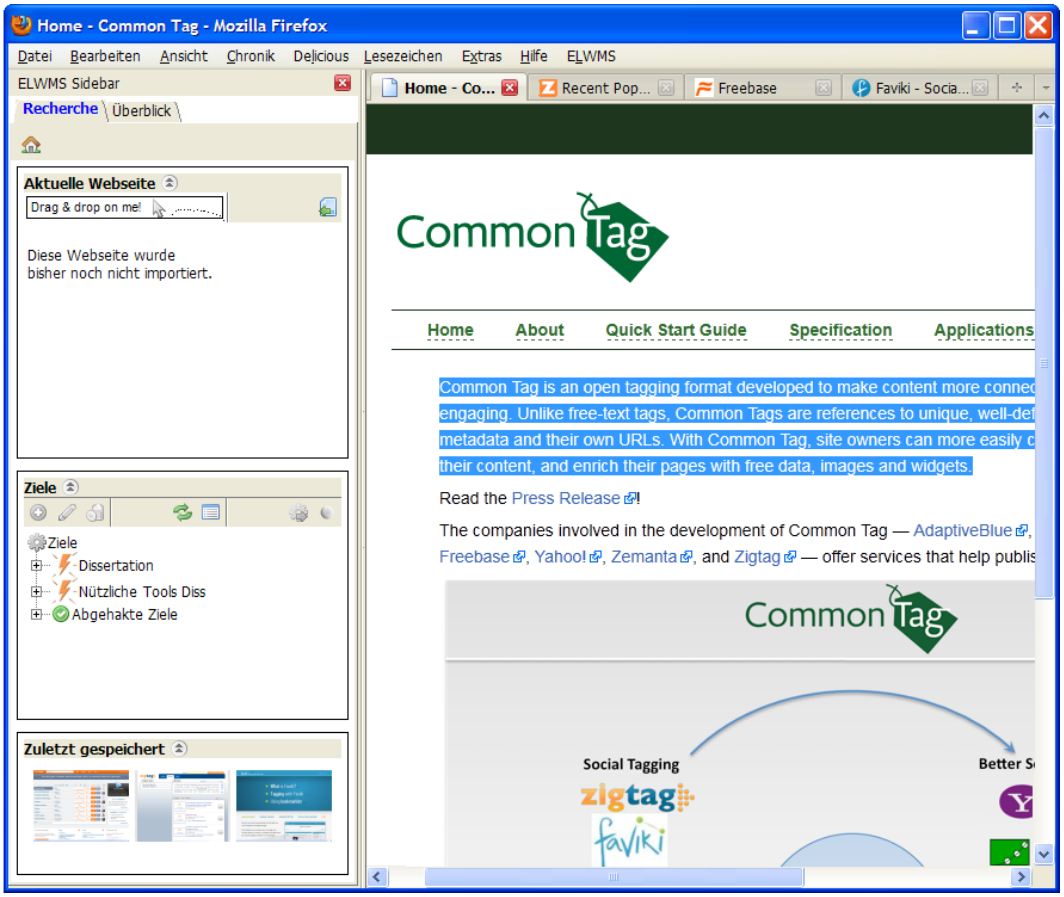


Figure 4.1: A selection from a web resource is to be saved with ELWMS.KOM by dragging the selected web resource fragment and dropping it on the sidebar.

In RBL, however, LRs (which are primarily web resources in the scenario of this thesis) are usually only authored at creation–time. Therefore, granularity here is an aspect of LRs that is primarily relevant for the differentiation of the parts of LRs that are matching a current information need and parts of LRs that are irrelevant. ELWMS.KOM supports storing content in the granularity of snippets by allowing to persist only a selected fragment of a web resource (cf. figure 4.1). This has the advantage that only the content that is currently needed is stored in the knowledge network and the learner does not need to scour the complete web resource for re–finding the important information. The full web resource is still available via its URL. For example, when a learner needs to solve a programming problem, she is able to only collect the fine–granular snippets that are relevant for her current task. Later on, she will be able to retrieve exactly the relevant part of the web resource that contains the solution from the knowledge

network. If she needs to see the snippet in its original context, she is able to open the source URL of the snippet.

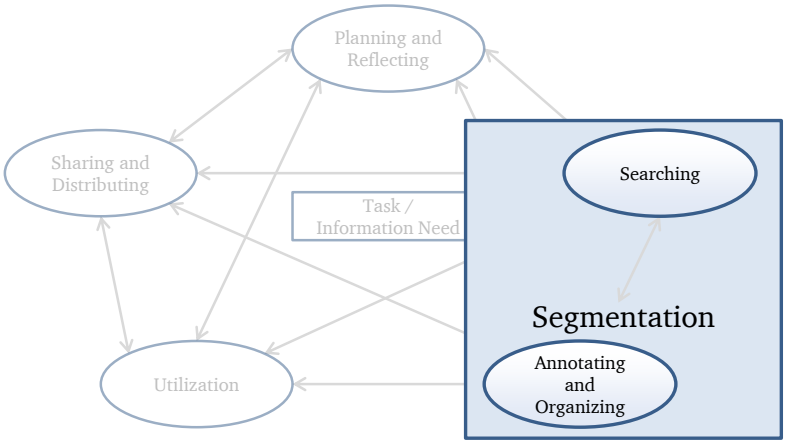


Figure 4.2: Automatic web resource segmentation supports retrieval and usability of organization in Resource–Based Learning

In the context of web resources, the notion of granularity therefore still is important. Hence, an approach to supporting the acquisition and target–oriented selection of relevant segments by providing *automatic segmentation* thus can support RBL (cf. figure 4.2). Although this functionality does not influence or support RBL directly, it can support usability in the process of acquiring relevant LR snippets. In its current state, ELWMS.KOM allows storing consecutive selected snippets of a web resource by manually marking, copying and pasting the content. However, often multiple, not necessarily consecutive key passages are considered as relevant. In the current version of ELWMS.KOM, learners have to create either different resources in the knowledge network or have to manually mark, copy and paste text passages several times. A solution to this could be an automatic segmentation of a web resource and allowing the learner to select the segments that she deems relevant in one step, eliminating the effort of having to gather all relevant passages manually. Another use case is checking whether the selected segments have been updated in the meantime. Especially in collaboratively edited wikis, text is prone to be edited often, possibly adding important content that might be relevant for a learner. Thus, learners could benefit from an approach to automatically detect whether a relevant passage of a web resource has been altered.

Further, an automatic segmentation of web resources constitutes a base technology that can be employed in different IR and usability settings, e.g. by enabling segment–based retrieval [48], filtering of unwanted content like advertisements [49], enabling small–screen browsing [9, 110] or segment type classification [61].

4.1 Introduction

As LRs consist of a whole web resource or parts thereof, the web resource's structure has to be taken into account for examining the concept of granularity. For example, a typical web page does not only consist of the actual information of interest (the *content*) but rather encompasses many other components like a navigational structure, advertisements, copyright information, headers and footers and, as mashups[1]

[1] In web development, a mashup is a web page that combines data or reusable parts from two or more sources on the Web to create a new service, aggregation, presentation or functionality [122].

have become a popular technique on web pages, so–called *widgets* that provide a functionality that can be embedded. Further, it is also common for web pages to contain multiple topics that do not necessarily belong together. Thus, the separation of the different content types has been a well–known issue in IR. Segmenting a web resource into *coherent* blocks can address this concoction of content, presentation and navigation by allowing retrieval algorithms to ignore blocks as navigation and advertisement or treating these blocks individually [44].

4.1.1 Coherent Segments of Web Resources

Coherence in this regard is based on three principles introduced by Bar–Yossef and Rajagopalan that are commonly called the *hypertext IR principles* [14]:

The relevant linkage principle states that a resource links only to another *relevant* web resource. This principle is very important and, for example, provides the foundation for algorithms that web search engines incorporate (e.g. HITS [103]).

The topical unity principle states that co–cited documents are related. This is again significant for web search engines.

The lexical affinity principle states that the proximity of text and links gives a clue about the relevance of text and linked resources. This means that the proximity of text and links can represent linked pages appropriately, PageRank [147], for instance, is based on this principle. Thus, this implies that there is a certain block granularity in a web resource that represents a meaningful, topically coherent segment.

These hypertext IR principles primarily apply to the interlinked structure of the web as such and do not specifically note the granularity of single web pages. However, the lexical affinity principle states that there is the concept of *proximity* that is important, implying that there are different areas in a web page that do not necessarily cover the same topics. This notion of proximity can be generalized to the concept of *coherence* that encompasses following properties and observations:

1. A coherent segment of a web resource is topically consistent. For texts, it is common practice to denote a change of topic by structuring texts visually, e.g. by grouping text into paragraphs or partitioning a text with headings. This is reflected in the visual appearance of text in rendered web resources.
2. There are logical entities in web pages that denote atomic communication acts, e.g. there are comments in a blog post that originate from different authors. This is reflected in the layout of a web resource.
3. In general, it is considered as good user interface design to visually group functionalities that are semantically similar[2]. For example, all links that belong to the navigation of a web page should be grouped in a specific location on a web page, usually the *navigation menu*.
4. The hierarchical structure of a HTML page often correlates with the topical, logical or functional relatedness of the content on the same level. For instance, in typical web pages, all information that can be considered the main content (e.g. the blog post on a blog page) is located in a sub–tree of one single HTML element [82]. Therefore, this hierarchy may hint to different "parts" of a page and is important to be taken into account.

[2] `http://www.readwriteweb.com/hack/2010/09/6-tips-for-building-coherent-s.php`, *retrieved 2010-11-15*

4.1.2 Structure of this Chapter

This chapter describes an approach to automatically segment a web resource into coherent fragments called Hybrid Recursive Segmentation Approach (HYRECA). Section 4.2 represents a brief excursion introducing some basic paradigms and concepts of HTML and the Document Object Model (DOM). Section 4.3 presents an overview of related work in this field of research. Then, section 4.4 introduces the design goals of a hybrid visual–structural approach, presenting an algorithm that takes these design goals into account and attempts a hierarchical segmentation of a given web resource. This is followed by an evaluation of the approach including selected findings in section 4.5. Eventually, section 4.6 gives a conclusion and establishes further perspectives.

4.2 HTML and the Document Object Model — a Short Summary

The "lingua franca" of the WWW is HTML — a markup language that is a logical representation of a web page's content. The current standard recommended by the World Wide Web Consortium (W3C), HTML4, is based on the Standard Generalized Markup Language (SGML). SGML is a predecessor of the Extensible Markup Language (XML) that does not yet have to be well–formed. As XML enforces strict nesting, it is less error–prone to parse. To accommodate this, there is a second W3C recommendation called Extensible HyperText Markup Language (XHTML) 1.0. The latter is — according to [165] and *builtWith*[3] — becoming more prevalent on the WWW[4], although the successor to HTML, HTML 5, will not be based on XML. In the following, HTML and XHTML are mentioned synonymously if not stated otherwise. HTML is primarily targeted at representing documents for display in a *web browser*. HTML pages are parsed into a tree–like structure of *element nodes*, which is the computer's internal representation of a HTML document (see figure 4.3 for an example). Usually, access to the element nodes that contain the textual information contained in the web resource is provided by an Application Programming Interface (API) called the DOM. HTML documents have exactly one *root node*, the `html` element. Each DOM node can have attributes and child nodes, thus a tree structure is built from the root node. For more information, especially the nomenclature of the relation degrees of DOM nodes, see the respective W3C standards [185, 154, 127].

Further, elements can be differentiated based on their flow mode in the rendered HTML page. *Inline elements* are rendered in the normal text flow like a single character, building a continuous line and only wrapping to next the line, if the remaining horizontal space is too small (e.g. the elements `em` or `span`). *Block level elements*, however, define an own block scope and break the current flow of elements. For example, headings (`h1-6`) are rendered on their own line, breaking the preceding flow of elements. A listing of the HTML elements and their respective flow mode can be seen in table B.1 in appendix B.

It is common practice to separate content and layout of a web resource by using CSS for styling HTML. Further, there is a trend to make HTML more "semantic" by sticking to certain best practices, e.g. giving nodes appropriate IDs or class names (e.g. adding the ID `content` to the `div` that contains the main textual body). Further, there have been efforts in establishing *microformats* to semantically markup entities like social connections (XFN[5]), contacts (hCard[6]) or events (hCalendar[7]) in HTML. This is often

[3] `http://trends.builtwith.com/docinfo`, *retrieved 2010-10-01*

[4] From 2006 to 2010, the ratio of HTML 4 pages (*transitional* and *strict*) has decreased from 70% to about 23%, whereas the ratio of XHTML 1.0 has increased from 15% to about 46%.

[5] `http://microformats.org/wiki/XFN`, *retrieved 2010-11-02*

[6] `http://microformats.org/wiki/hcard`, *retrieved 2010-11-02*

[7] `http://microformats.org/wiki/hcalendar`, *retrieved 2010-11-02*

```
<html>
  <head>
  <title>A page</title>
  </head>
  <body>
    <h1>Title</h1>
    <p>This is
    <a href='…'>
      a link
    </a>!
  </body>
</html>
```

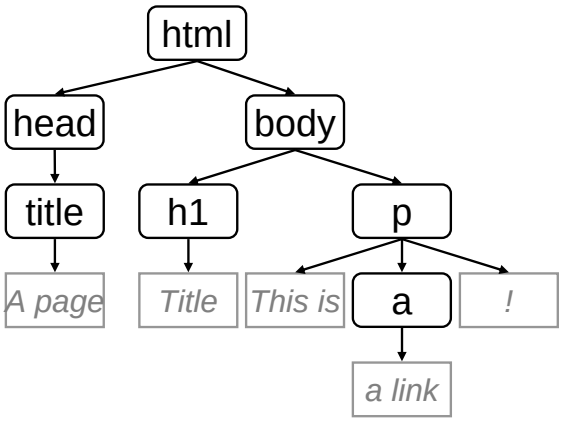


Figure 4.3: Example of a Document Object Model tree structure built from a simple HTML document. The attribute of the a element is not depicted.

seen as a necessary step towards the semantic web that provides the means to search in, interweave or infer over *linked data*[8]. HTML5 will provide better support of semantic markup, but it is still in draft status [88] and therefore usage in productive environments is not yet recommended.

4.3 Approaches to Segmenting Web Resources

In this section, the *page segmentation problem* is formally defined and related approaches are discussed.

4.3.1 A Definition of Coherent Segments

As described in section 4.1, web pages usually consist of structurally independent, coherent blocks that serve different functionalities. Over the last decade, many researchers have addressed retrieval and presentation that accommodates to this smaller level of granularity, for instance by the content source [46] in IR tasks or adapting web pages to small screen displays [9, 110]. Over the years, several different terminologies like "web component" [34], "web page block" [179], "semantic block" [44] and "clippings" [121] have emerged. In this chapter, the term *segment* is used to describe these fragments of web resources.

Cai et al. [44] see *coherence* as an important attribute of a segment. They state that the *semantics* of a segment define the coherence. However, they do not elaborate their definition and eventually reduce coherence to visual coherence as it is perceived by humans. In this thesis, the concept of coherence spans functionality, topic and structure (cf. section 4.1) of a segment. Thus, a definition of a *coherent segment* is given as follows:

Definition *A* coherent segment *is part of a web resource that obeys following constraints:*

- *A segment should only serve a single* functionality*,* communication act *or* topic*. For example, a valid segment could encompass all navigation elements, like a menu (functional aspect) or contain a single user's comment to a blog post (communication aspect).*

[8] `http://linkeddata.org/`*, retrieved 2010-11-04*

- *A segment must not be nested within another part of the web resource that serves exactly the same functionality or topic. For example, a part of a comment may not be considered to be a valid segment, rather the whole comment should be seen as a segment.*
- *A segment may contain other segments. For example, a segment containing a blog post may be seen as a functional segment (serving the functionality to publish a text), whereas different sub–parts like different paragraphs can be seen as topical segments (containing the flow of argumentation).*

The *web resource segmentation problem* is to divide a page into a set of such coherent segments. Without the second constraint from this definition, the granularity of each segment corresponding to exactly one DOM node would be a valid solution. The third constraint serves to reflect the different abstractions of the notion of “functionality” and “topic”. For example, a web resource representing a forum page can be seen as a unit regarding the content of the discussion, but, on a more fine–granular viewpoint, each post can be considered a segment as well, as it serves the functionality to add a comment to the overall discussion (see figure 4.4). Hence, a segmentation process can respect these granularities by returning a hierarchical view on the segments.

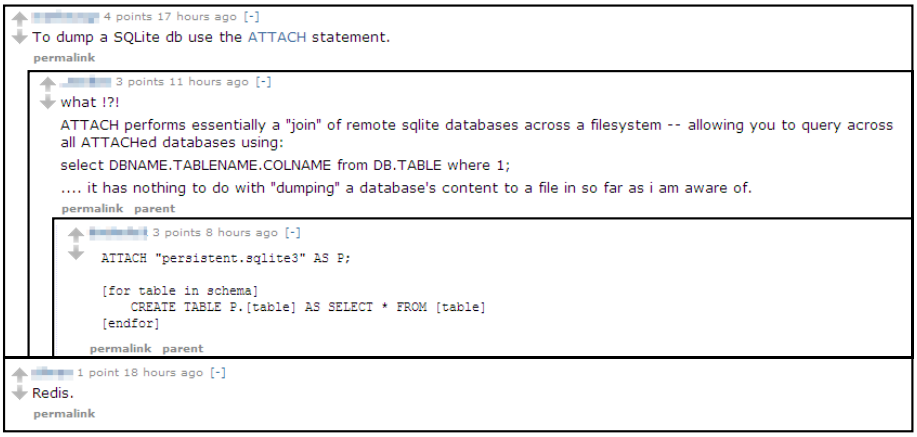


Figure 4.4: Example of nested comments on a Reddit community page. The hierarchical structure is highlighted with boxes.

In accordance with related work, this thesis assumes that all of the information contained in a web resource is to be segmented, hence the web resource should be completely covered by the segments. *Filtering* irrelevant segments or detecting informative sections has been independently researched by several authors, e.g. [119, 57]. Thus, segments that are not considered content are retained in the results.

4.3.2 Related Work

The different approaches to segment web resources can be grouped into three different categories that take into account different types of information. First, there are approaches that take statistical properties of the HTML markup into account, namely the density of HTML elements versus the plain text content of a web resource. Further, there are approaches that build a DOM representation of the web page’s HTML and analyse the resulting DOM node tree. Eventually, there are approaches that render a visual representation of a web resource (the same way a web browser does) and analyse the resource based on layout properties like whitespace. Further, there are hybrid approaches that take into account more than one representation.

Statistical Approaches

Champandard [49] presents an approach to extract relevant text from web resources using statistical properties of the ratio between a page's markup and the content. Using this information, this approach is able to differentiate whether a HTML line is part of the actual relevant content of the page. With the application scenario of filtering irrelevant information, this approach considers content not being part of the main textual part of a web page as *noise*, e.g. navigation, headers, footers and copyright notices. Champandard trains a neural network with a manually labelled training set which handles the classification into noise and informative sections. As features, he utilizes the *density* of the HTML, which is the ratio between textual content and HTML markup. Further, the byte counts for text and HTML markup are taken as features. Champandard claims good results, minimizing the false positive / false negative error compared to a naïve approach not based on density by 80%.

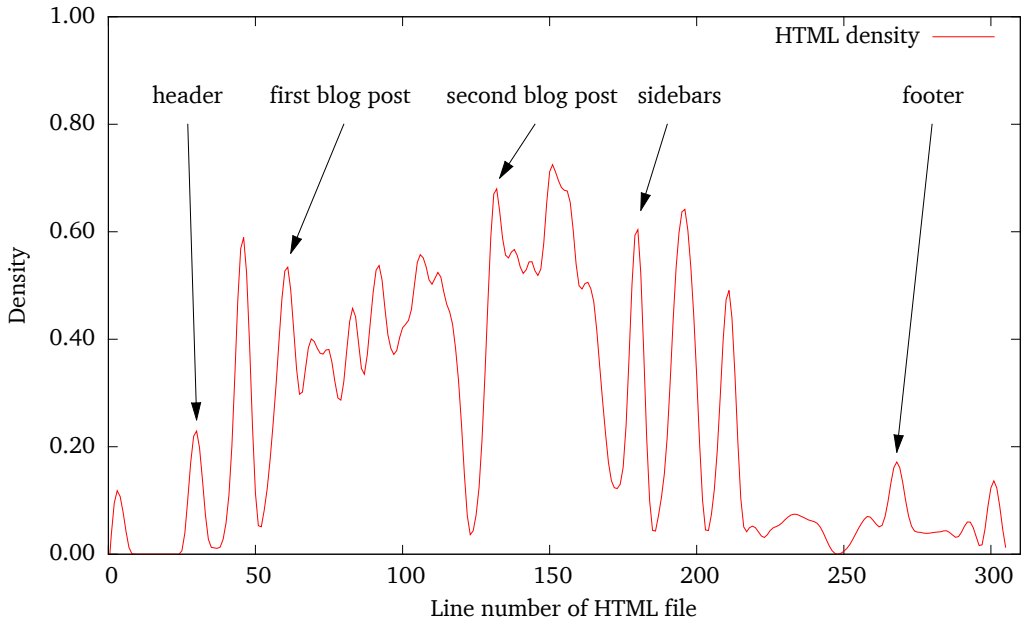


Figure 4.5: Plot showing the smoothed density of an exemplary blog start page. The "dent" in the graph between the two blog posts is caused by the HTML markup of the underlying template, wrapping the blog posts.

This approach has some issues, with the granularity level of the segments being the most severe one: as Champandard calculates the HTML density line per line and does not take into account the hierarchical structure of the HTML, the approach is not able to cope with very dense or very light markup. Dense markup occurs where markup is not "pretty–printed", e.g. there are no line breaks between elements. As line breaks do not affect the representation of the web page, some web applications filter them out in order to save on bandwidth. Very light markup occurs where the HTML markup is indented for each element, here a density–based approach would favour short textual sections like navigation texts or copyright notices. Further, the resulting segments are generally not well–formed HTML and structural information of dense segments in one page is lost [74]. Eventually, this approach does not take into account the connectedness of segments, exposing untypical spikes like shown in figure 4.5 like in the header that have to be handled appropriately. However, Champandard claims to achieve good results in realistic settings [49].

DOM–based Approaches

There are several approaches that do not work on the level of the HTML page's markup but rather operate on the hierarchical node tree representation of the DOM. As many web pages are nowadays generated using some templating mechanism, this is a property that most of the DOM–based approaches utilize. A *template* of a web page is given if there is a common layout frame for several web pages from the same host. This template is filled with the textual content by a *template engine* in the generating web application. For example, a blog application commonly has a template denoting the layout frame for different page types. While the blog start page aggregates the *n* newest blog posts, a blog post page displays one post and comments. This structure of the two pages are pre–defined in templates (so–called *themes*) and can usually be chosen by the owner of the blog. Often, noise elements like header, footer, navigation and advertising are part of that template. Gibson et al. [80] estimate that 40%–50% of the total content in the web is template code. Thus, *template detection* is an important field of research in web mining.

Most approaches detect templates by taking samples of different pages from the same website or by taking samples of one web page at different points of time[9]. For generating a structure representing a template, there are two challenges that have to be addressed:

1. The web page has to be divided into blocks.
2. The frequencies of each block in the set of pages originating from the same template have to be determined.

The first step — dividing the web page into blocks — is usually achieved by using heuristics. These heuristics are often dependent on the best practices and the state of technology of their time. For example, in the days before CSS2 styling became supported by all major browsers, complex web pages were often designed using the `table` tag, thus several authors [119, 201, 53] proposed to examine only the `table` tag and its respective child tags, `td`, `tr`, etc. However, nowadays using tables for layout purposes is discouraged for several reasons, mainly accessibility [32], and thus these approaches will not work with modern web pages that rarely use table based layout. Most newer approaches [58, 107] therefore define sets of HTML elements that serve as valid block containers in absence of table based layout. For example, Debnath et al. [58] use the elements `table`, `tr`, `p`, `hr`, `ul`, `div` and `span` for denoting block containers or block separators. A recursive partitioning process is executed until none of these elements is contained in any block. However, these blocks tend to be very fine–granular (e.g. due to usage of the element `p` representing a textual paragraph as a separator), and therefore do not adhere to the definition of segments given in section 4.3.1.

The second step — determining the block frequencies for the different pages — yields the common blocks that are contained in the different pages, and thus denote a part of the underlying template. However, as small differences may occur even in template blocks, e.g. the numbers of a visitor counter, this step has to provide a fuzzy comparison. There are multiple different similarity measures used for this fuzzy comparison. Ramaswamy et al. [156] use the shingling algorithm to generate fingerprints of each block that only change little if the block content changes little. Miloi [136] evaluates the Levenshtein Distance [117] and a simple distance measure based on term counts. Yi et al. [201] only use the internal tag structure of a block and ignore the textual content.

These DOM–based approaches are usually very efficient and yield good results, especially when more than two samples of the same page layout are available and a more reliable content structure can be

[9] This will only work with pages that change their content frequently, therefore the first option is usually preferred.

inferred [201, 156, 57]. However, in settings that do not guarantee to provide different instances of the same template, these approaches are not applicable.

Visual Approaches

Visual approaches utilize visual layout information of a web page and thus simulate the human perception of a web page to some extent. The main idea of visual approaches is that web pages are designed to *present* information to humans and therefore follow certain layout principles and best practices. Cai et al. [44] present a segmentation algorithm called Vision–Based Page Segmentation (VIPS) that analyses visual separators like whitespace or block sizes in the page, using this information as an indicator for distinguishing different visual segments in a page. They state that these visual segments correspond to *semantic segments*, clustering different functionalities and topics very well. VIPS has been applied to different IR settings [45, 42].

VIPS consists of three steps:

1. The page is divided into blocks. Each node in the DOM tree is recursively traversed and is tested for comprising a block by use of heuristics. These heuristics decide based on node name, background colour differences, the size of the node's sub–tree in pixels and textual cues whether the current node forms a block. For example, if one of the node's children has another background colour, this is a hint to segment the node further. Further, the size heuristic prevents further division of a node when its rendered representation is smaller than a predefined threshold. Each block that has been identified is assigned a measure of coherence, called Degree of Coherence (DoC), that ranges from one to ten and takes into account visual cues and structural information. Its value is determined by rules, e.g. if all child nodes of a DOM node are text nodes or inline nodes and have the same font weights and styles, the DoC is set to 10. If the font weights and styles differ, it is set to 9.
2. The blocks that have been identified in the first step are used to generate separators. Starting with one separator that covers the whole page, three rules are applied to each block and each separator: (i) If the block is contained in the separator, split the separator. (ii) If the block overlaps with the separator, update the separator's parameters size. (iii) If the block completely covers the separator, remove the separator. This process is illustrated in figure 4.6 for horizontal separators. Separators are weighted based on layout and structural properties of the adjacent blocks.
3. Adjacent blocks are recursively merged up to the separator with the maximum weight, starting with the lowest weighted separator. The merged block's DoC is updated with the maximum weight of the adjacent separators.

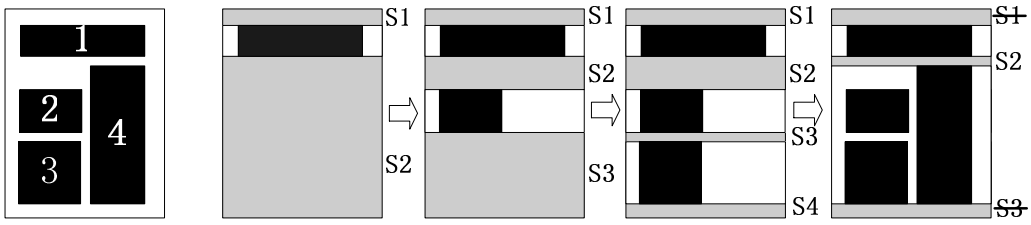


Figure 4.6: Example for Horizontal Segmentation using VIPS (cf. [44])

VIPS terminates if all block's DoC is greater than a predefined threshold, and merged blocks are returned as segments. If a block does not meet the DoC threshold, the algorithm is started again with this block as input.

VIPS was evaluated by five participants judging the results of 600 web pages [44]. The participants were asked to value the results for each page as “perfect”, “satisfactory”, “fair” or “bad”, which are very subjective classes. Cai et al. claim that 93% of the web pages were either labelled “perfect” or “satisfactory”. Further, Cai et al. evaluated the quality of VIPS based on the improvement of a web IR task, namely *query expansion*[10]. They state that using VIPS, selecting only the relevant segments and not the whole web page for detecting expansion terms, could improve the query expansion task by 22%.

Hybrid Approaches Combining DOM and Visual Analysis

Xiang et al. [198] present a hybrid approach, called Pattern Analysis and visual Separators (PAS), combining DOM–based pattern analysis and visual analysis based on VIPS. They state that humans, when dividing a web page into segments, are strongly guided by visual patterns, i.e. repeating visual structures. Therefore, they add a detection of repeating patterns to the consideration of visual separators. Repeating patterns are existent in a lot of web page’s templates, e.g. as comments in blog posts as displayed in figure 4.7, that share a discerning structural similarity.

Figure 4.7: Example of repeating patterns in web pages, here comments in a weblog.

The algorithm consists of three steps:

1. First, the nodes in the DOM tree are enriched by adding information about their visual properties, e.g. their position on the web page or — in case of text elements, their font style and weight. Then, adjacent inline elements, i.e. elements that are rendered in the same line as the preceding elements (cf. section 4.2), are merged and visual separators like the `hr` (defining a horizontal ruler) are removed from the DOM tree.
2. Then, patterns in the tree structure are detected and the respective nodes in a pattern are grouped. The greedy algorithm that Xiang et al. employ searches for repeating patterns below each parent node. Patterns may encompass multiple sibling nodes and may have separating nodes between patterns.

[10] The idea of query expansion is to find more relevant documents for a given query by expanding the query, i.e. automatically adding additional related search terms, e.g. synonyms.

3. Eventually, detected patterns are wrapped into a new `pattern` element that is inserted into the DOM tree, effectively breaking the validity of the DOM. Patterns are grouped into a `group` element. Nodes that are not children of `group` nodes are analysed by the VIPS algorithm.

Xiang et al. evaluate their approach by having five participants manually segmenting 40 web pages from 13 web sites. These resulting segmentation is compared to the results of PAS and VIPS. Xiang et al. show that the performance of this hybrid approach matches VIPS for a large granularity of segments but significantly outperforms the original VIPS results for a small granularity.

4.3.3 Discussion of Related Work

The presented classes of web resource segmentation approaches, ranging from the statistical to hybrid visual and DOM–based approaches, increasingly add a level of sophistication. As they are designed to serve a specific task or goal, some approaches are not necessarily applicable to a general page segmentation task and have limitations. For example, the statistical approach is entirely based on lines and therefore lacks the means to cleanly discern between segments, as it employs a fuzzy rule to differentiate between noise and informative content. The result of this approach is not a set of multiple segments but rather a coarse segmentation of the HTML source code based on lines. The DOM–based approaches are better suited to these requirements: the results are — depending on the approach — fine–granular based on single DOM nodes, but they completely disregard visual information that allows emulating the perception of humans. The visual approach takes this information into account but does not harness from the observations that repeating patterns provide good segmentation cues. The hybrid approach focuses on those patterns, but still has the limitation of breaking the validity of the analysed page's HTML and finding only patterns that are contained in one DOM node. Therefore, this approach could not detect repeating pattern that are hierarchically grouped, like e.g. hierarchical comments in blog posts (cf. figure 4.4). Further, Xiang et al. 's algorithm requires an explicit specification of the expected segment size. In non–supervised setting, this is not feasible, a generic approach should recognize small and large segments automatically.

Thus, in the following section a novel approach to segmenting web resources is proposed that accounts for the short–comings identified in related work. Especially the validity of the analysed resource's HTML is of concern, as it makes the resulting fragments usable in scenarios that expect to get valid HTML as input, e.g. small–screen rendering or displaying segments for a user selection in ELWMS.KOM.

4.4 HYRECA — A Hybrid, Hierarchical Approach to Web Resource Segmentation

In this section, an overview of a novel approach called HYRECA is given that incorporates the DOM–based as well as a visual analysis and the detection of repeating patterns. Based on the short–comings of related work and requirements stated in section 4.3.1 HYRECA has following design goals:

1. Providing segments that accord to the coherence principles stated in section 4.3.1 in different granularities. Segments should be able to nest, as there are large coherent segments that encompass multiple smaller coherent segments (e.g. the part of the blog post page containing all comments).
2. Covering the whole page in segments. There should be no parts of a web resource that are not contained in a segment. Further, the granularity of the respective segments should be automatically derived from the features of a web resource.

3. Working with a wide variety of web pages. Therefore, HYRECA should be independent of a certain structure or design of a web page. Other approaches rely heavily on a certain page structure (e.g. using tables) and therefore are not compatible with modern paradigms of web design.
4. Exploiting diverse information like the DOM and visual representations of a web page.
5. Taking into account repeating patterns, even if they are not only contained in one DOM node (in contrast to [198]).
6. Not breaking the validity of the analysed web page. It should still be viewable with common web browsers and thus can be applied to different scenarios like small screen browsing or segment–wise IR. This design goal is violated by most related approaches (e.g. [44, 198]) which makes them inapplicable to scenarios where the segment detection is only an intermediate step followed by consumption in a web browser or similar application that expects well–formed HTML like ELWMS.KOM.

In this section, an algorithm is described that meets these design goals and provides a segmentation of arbitrary web resources.

4.4.1 Description of HYRECA

HYRECA consists of five steps that are executed in succession (see figure 4.8). Preliminary tests showed that these steps yield results with a different reliability. For example, the pattern finding step returns very reliable results, whereas the visual segmentation sometimes yields inconclusive segments. Further, as the class and id heuristics are aimed at detecting very coarse granular segments that are anyway expected to be found by the visual analysis, they add value only in special cases (e.g. when the visual analysis fails due to inaccurate rendering caused by the used HTML rendering engine). Thus, the order of the steps is important with regard to the reliability of the results. Additionally, the separate steps can access the results of the preceding steps, so that the heuristics can accord for potential errors in the first steps.

Pre–processing The web page's HTML and linked CSS files are downloaded and parsed, resulting in a DOM representation of the web page (cf. figure 4.8a). Then, the page is rendered[11] and the DOM's nodes are enriched with visual cues like background colour, dimensions and location of the node. In this step no segment candidates are selected.

Pattern Finding This step identifies recurring patterns in the DOM structure. This pattern detection is based on the assumption that patterned segments are not only structurally similar but visually similar as well (cf. figure 4.8b), and thus are recognized as segments by humans. This step will be elaborated in section 4.4.2.

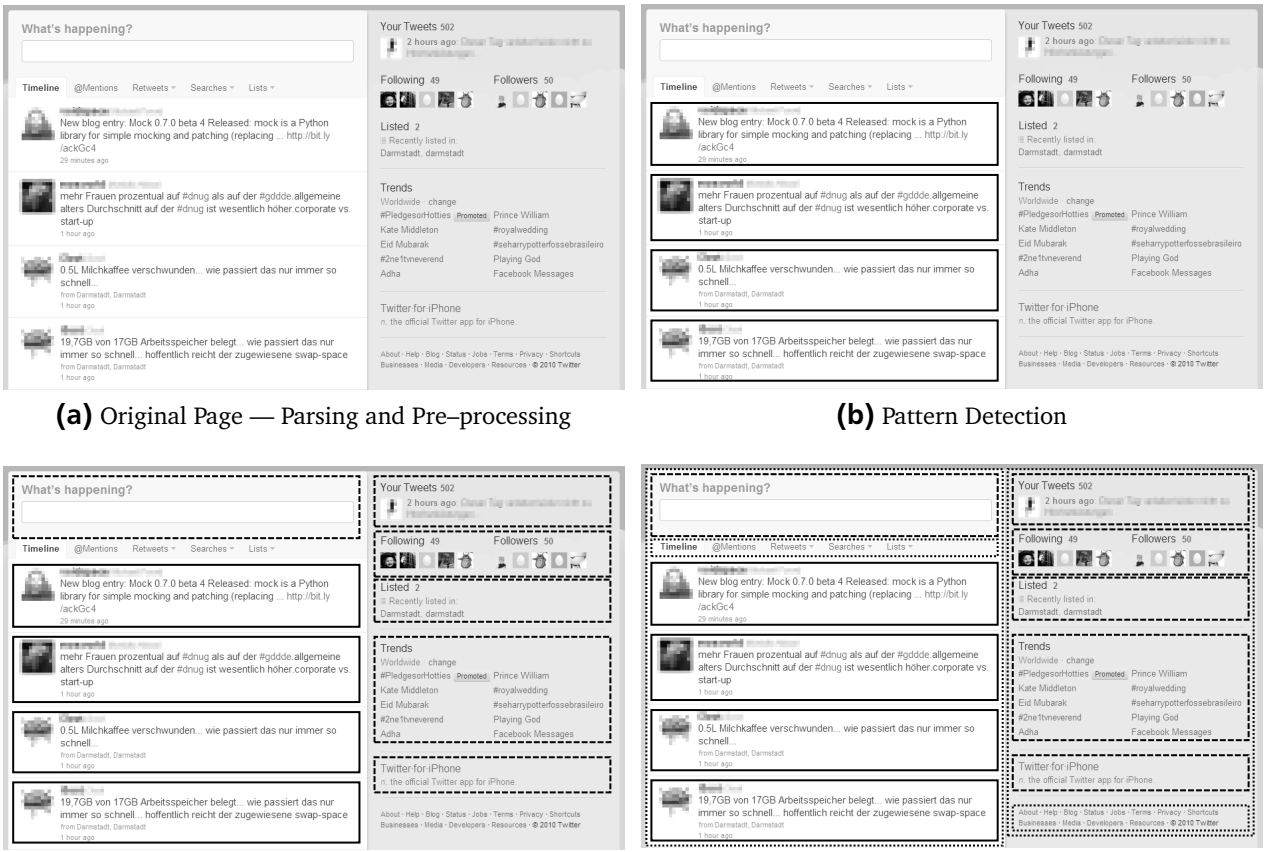
Visual Analysis The visual cues added in the pre–processing step are the basis of this step. It emulates the simulation of the human perception to some extent by applying heuristic rules that result in a grouping of elements as humans would group them (cf. figure 4.8c). These rules only work on visual parameters and ignore the content of the elements. Details are given in section 4.4.3.

Class and ID heuristics In HTML, the `id` and `class` attributes are used for applying CSS styles and providing hooks for JavaScript. Allsop et al. [1] and Google [82] performed an extensive analysis on how these attributes are used. Both come to the conclusion that there are values that are very common, denoting "best practices" for web designers on how to name functional and representational blocks of HTML. For example, common names for these attribute include `header`, `footer`, `content`, `sidebar` or `menu`. HYRECA identifies the elements with the most common attributes as

[11] Using the Cobra Rendering Engine, `http://lobobrowser.org/cobra.jsp`, *retrieved 2010-11-05*.

segments. Usually, these segments are in a high hierarchy level and reflect the basic skeleton of a web page (cf. figure 4.8d).

Post–processing As the preceding steps generate a lot of segment candidates of which some are redundant or superfluous, the post–processing is needed to clean these results. For example, often nested segment candidates are detected where the outer segment does not add any displayed content (i.e. textual content or images) to the inner segment. In this case, the outer segment candidate is discarded. Further, to consider a segment to be valid, it has to contain a minimum of textual content and have a minimum size. Segments not cohering to these requirements are discarded.



(a) Original Page — Parsing and Pre–processing

(b) Pattern Detection

(c) Segmentation using Visual Hints

(d) ID / Class Segmenter

Figure 4.8: The process steps of HYRECA visualized in an exemplary segmentation of the Twitter home page. The Post–processing step is not depicted here, as it merely cleans up superfluous segment candidates.

Marking and storing segments is performed between each process step. Thus, the next process steps can access the segmentation results of the preceding steps. For each found segment or group, a unique class name is generated with the prefixes `HYRECA-` (for segment root nodes) or `GROUP-` (for sibling nodes that are the root of a pattern occurrence) and applied to the respective node. Thus, the structure and therefore the validity of the DOM document stays intact and no further nodes are introduced (in contrast to [198]).

In the following sections, the processing steps of pattern detection and visual analysis are refined.

4.4.2 Pattern Finding

Xiang et al. [198] state that users are highly guided by repeating patterns when viewing a web page. In web applications that generate pages like weblogs, wikis and forums, these patterns are often rendered by a templating engine that embeds different entities of content (e.g. comments in a blog post) in a common markup fragment. Thus, they share a similar HTML structure. However, not all patterns in a web page are automatically valid segments. For example, lists in HTML are marked up using `li` elements for each list item. According to the definition of coherence given in section 4.1, these single list items will not be considered a valid segment. Therefore, a certain complexity of a pattern's markup structure has to be assured.

Further, not all occurrences of a pattern have to have exactly the same structure. For example, a blog comment containing additional markup, e.g. a link or an image, should still match the overall pattern. Thus, the matching process should be fuzzy, i.e. being able to abstract from small discrepancies from the pattern.

Building the Node Fingerprints

For finding repeating patterns, each DOM node has to be available in a form that represents its complete sub–structure, i.e. all direct and indirect children. This representation is called the *fingerprint* of the node. It is built by recursively traversing all child nodes of the respective node in *preorder*[12]. For each node, the name of the node is appended to the fingerprint. In order to account for small discrepancies between the different occurrences of a pattern, inline elements are ignored. These inline elements are usually used to format text and do not contribute to a distinctive representation of the *structure* of a DOM sub–tree. Paragraphs and links (the tags `p` and `a`) are treated specially. Although `p` is a block level element, it is ignored, as text segments may contain multiple paragraphs. Links, on the other hand, have a functional aspect as they allow displaying a hyperlink that can be used for navigating, and therefore the tag `a` has to be represented in the fingerprint. For example, the structure of blog comments usually contains a link to the comment's author's homepage, so this is an important feature as it represents the functionality to link to the author's blog. However, if an author writes a comment including many links, these links are part of the content of the comment and not of the comment structure and therefore should be ignored. This conflict is solved in this approach by contracting multiple links that are siblings into one meta–link `a*`. Hence, the occurrence of many links in one node does not have a large impact on the structural fingerprint of this block. An example for the fingerprinting result including a contraction of multiple `a` tags is displayed in figure 4.9.

Matching similar Pattern Candidates

After a fingerprint has been calculated for each node, the second task is to find repeating occurrences of the fingerprints.

Definition *In order to decrease the complexity of the algorithm, a pattern has to follow two constraints:*

1. *A pattern must occur at least twice below the same parent. Xiang et al. [198] also make this assumption, exceptions thereof are handled by the visual segmentation approach.*

[12] A preorder tree traversal is when the current node is handled before each child is handled recursively.

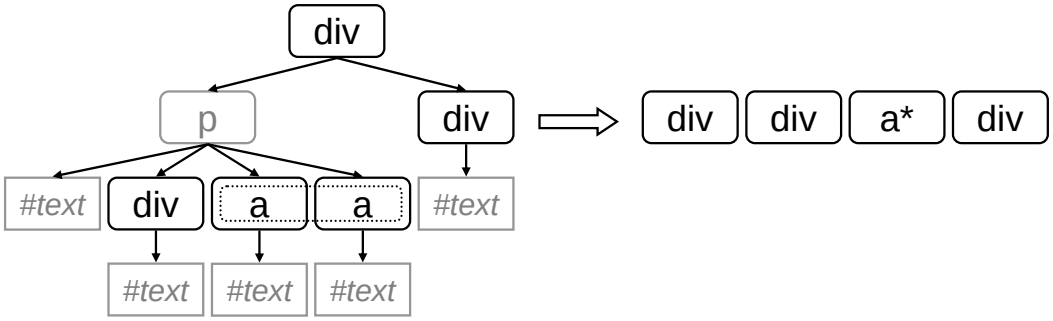


Figure 4.9: Example of fingerprint extraction from a DOM sub–tree. The DOM tree is traversed preorder, the nodes `#text` and `p` are ignored and multiple `a` successive nodes are contracted.

2. *Occurrences of patterns are consecutive. This means that if the pattern order is broken by unexpected non–patterned nodes, the respective pattern will not be detected.*

Recursively traversing the tree starting with the root node, patterns are searched within the child nodes of the current node. As the pattern may span multiple root nodes, the algorithm looks for repeating occurrences of the same order of node types as pattern candidates. Figure 4.10 shows an example of a pattern spanning two root nodes. The amount of pattern root nodes found for each pattern is called the *pattern length*. By taking into account the second assumption, the number of comparisons is dramatically reduced. If n is the amount of child nodes in the currently examined node, and l is the pattern length, $O((\frac{n}{l})^2)$ comparisons would be needed without the assumption, whereas with the assumption this is reduced to the complexity $O(\frac{n}{l})$. The actual comparison is based on the Levenshtein distance [117] (a metric providing the edit distance) of the fingerprints, matching patterns despite smaller aberrations in the fingerprints. Two fingerprints are considered similar, if the normalized Levenshtein distance, i.e. the ratio of edit operations to the fingerprint length, is below a certain threshold. Based on manual reviews and empirical results, this threshold has been set to 85%.

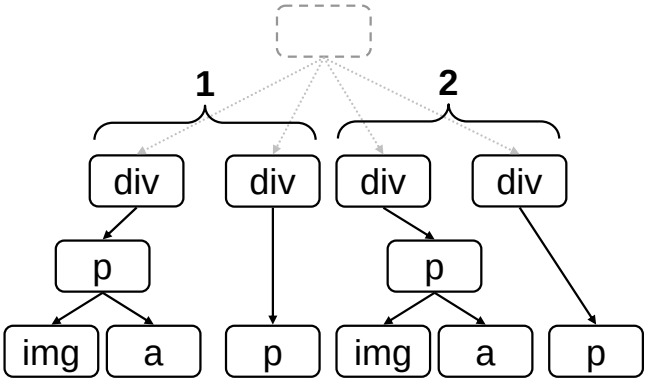


Figure 4.10: Two instances of an example pattern consisting of two neighbouring DOM sub–trees

There are further refinements to the selection process that base on heuristics. Thus, only certain elements (consisting of block level elements, e.g. `div`, `h1` to `h6`, `li`, `table` and `td`) are allowed to be the root node of a pattern in order to avoid misaligned patterns or false positives. Yet, there are exceptions, e.g. the `pre` element (displaying pre–formatted text in a block level element) is always excluded as a pattern root. Another exception is the table row element `tr`, as table rows are always a structural pattern by design.

If a pattern candidate has been found, the respective DOM nodes are marked by adding a *marker class* to the node's `class` attribute.

4.4.3 Visual Analysis and Grouping

In contrast to VIPS [44] presented in section 4.3.2, HYRECA does not aim to place separators in a web page to be segmented. Rather, it aims to group elements together which are likely to form a unit in terms of representation or functionality. The cue whether or not two DOM nodes can be grouped into a segment is based on the visual properties of the nodes.

Consecutiveness in Source and Layout

The segmentation process employing visual properties of a web page assumes that a segment that is *coherent* also is *consecutive*, meaning elements that are part of a segment share a certain proximity.

Definition *In HYRECA, there are two different notions of* consecutiveness*:*

Consecutiveness in source *Two elements* a *and* b *are considered* consecutive in source *if the* a *is the* previous sibling *of* b *in the DOM tree; this implies that* a *and* b *are also consecutive in the HTML source code.*

Consecutiveness in layout *Two elements* a *and* b *are considered* consecutive in layout *if their bounding boxes (i.e. their rectangular edges) in the layout share one edge.*

Therefore, while two elements that are *consecutive in source* may or may not be *consecutive in layout*, the other direction is usually true. However, there exist two notable exceptions that are not covered by HYRECA, being limitations of the approach:

1. Two elements that are separated in the source HTML can be made consecutive in layout by using CSS, e.g. using the `position:absolute` directive. However, web pages rarely rely on this mechanism, and therefore this case can be ignored.
2. The cells of a column in a table are consecutive in layout while they are separated in the source HTML. Therefore, a segment consisting of multiple table cells in one column are not recognized by the visual segmentation algorithm as proposed here. However, this case has not been observed in the evaluation data, therefore it is ignored.

The consecutiveness definitions given above are relevant, as they imply that DOM nodes that are analysed share the same parent and are direct siblings (cf. the definition from section 4.4.2). This enables HYRECA to find consecutive patterns very efficiently.

Segment Candidates

The algorithm starts with a given DOM node and subsequently checks whether the next sibling of the node can be grouped with the actual node or it belongs to a new group. The decision is based on visual properties of the current node and the semantics of the element type of the node. The following set of rules is applied in the given order:

1. A `#text` node (i.e. text enclosed between elements) always belongs to the current group.

2. A node that has already been grouped by a previous segmentation step (cf. section 4.4.2) is not added to the current group. The current group is concluded and the process is continued with the next node after the already marked group.
3. An inline element node always belongs to the current group, as it does not break the horizontal flow of text.
4. Paragraphs are added to the current group. As `p`–nodes are used to format continuous text into paragraph blocks, they are considered to be too fine–granular for a segment.
5. The elements `h1` to `h6` and `hr` denote a break in the current web page and mark the boundaries for two segments. Thus, if one of these elements occurs, the old group is closed and a new group is created.
6. An arbitrary element may either start a new group or contribute to the current group. The visual information like position, size and match of background colour are used in order to decide whether the element belongs to the current group. If a DOM node's background colour does not match the current group, a new group is assumed. Further, visual consecutiveness is tested by comparing the positions of the current group and the new node.

Figure 4.11 shows an exemplary segmentation of the Heise start page[13] viewed in HYRECA's resource viewer. All segments are highlighted using boxes with dotted lines, the same colour denotes assignment to the same segment. It clearly shows the hierarchical segmentation of the web resource's body with separate segments for the sidebars and segments that span consecutive occurrences of `h3` and `p` elements for the article excerpts.

[13] `http://www.heise.de`, *retrieved 2008-05-27*

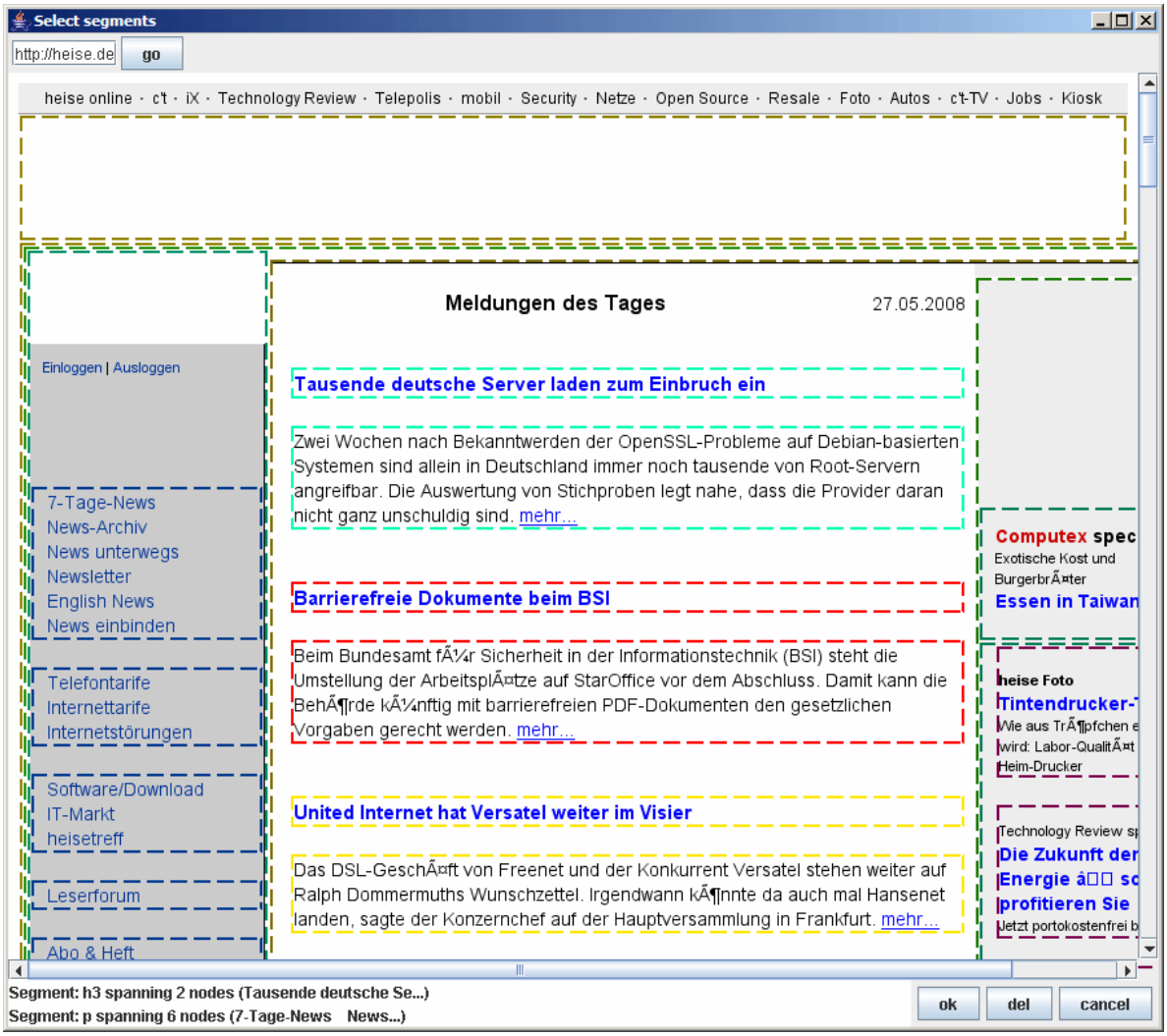


Figure 4.11: An exemplary segmentation of the Heise start page, displayed in HYRECA's resource viewer. Coherent segments are displayed in boxes using the same colour. Note that the header image and the advertisements are not displayed due to rendering problems introduced by the used rendering engine.

4.5 Evaluation and Results

The different approaches presented in section 4.3 present differing possibilities of evaluating the respective algorithms. However, they either do not publish their collection of test web pages (like [57, 198]) or they state web pages that are not existing anymore or have undergone a complete redesign (like [136]). Thus, it is not possible to achieve comparable data, making it necessary to create an own evaluation set.

Further, rating the segmentation just in categories "perfect", "satisfactory", "fair" and "bad" (as done in [44]) is not satisfactory, as such ratings are highly subjective and heavily depend on the ratio of errors and the length of a page. For example, ten segmentation errors in a short web page containing few segments are more significant than ten segmentation errors in a very long web page consisting of a lot of segments. Thus, an own methodology to evaluate HYRECA is presented in section 4.5.1. The evaluation design is explained in section 4.5.2, with the results being discussed in section 4.5.3.

4.5.1 Corpus Design

One design goal of HYRECA is that it should work without being limited to specific web pages or a certain web genre (i.e. a certain *type* of page like a blog post, a wiki page or an academic homepage). Therefore, a corpus that aims to serve as an evaluation foundation has to reflect these different types of web pages.

The evaluation corpus consists of 48 web pages manually assembled from five different categories of web pages (*Blogs*, *Company homepages*, *Web shops*, *News sites* and *Miscellaneous pages*). A full listing of the URLs can be found in appendix B.2. While this distribution originates on the idea to measure the quality and applicability of HYRECA on a broad spectrum of pages, it casts the research question whether the algorithm works equally well with the different genres. Thus, the following categories are taken into account:

Blogs This category is a set consisting of the main page and one single blog post page of the five most popular blogs from Technorati[14]. The blog posts contain a different number of comments.

Company homepages This category contains home pages of the ten biggest companies according to Fortune 500 in 2007[15] (e.g. BP, Walmart and Daimler).

Web shops This set consists of different web shop pages from eBay, Amazon, Dell and ASOS.

News sites This category contains pages from popular German and English news sites.

Miscellaneous pages This category is a collection of web pages that are interesting because they have certain properties that make them challenging for a segmentation approach. Included are wikis, forums, academic homepages and pages that are very lean on markup and heavily rely on visual representation.

The respective web pages were downloaded including all assets (e.g. linked images, advertisements, CSS and JavaScripts) and stored locally. The downloaded HTML was cleaned using Tidy[16] to prevent parsing and rendering errors due to invalid markup. HYRECA was executed and all found segments were visually highlighted. Two versions of segment visualization were provided, one highlighting the segments' outlines and one setting a unique background colour for each segment. Further, for each processed web page, the number of segments detected by the pattern and visual algorithm was stored. Four of the pages could not be processed by Cobra, the HTML rendering engine used, due to excessive

[14] `http://technorati.com/blogs/top100?type=faves`, *retrieved 2008-07-08*

[15] `http://money.cnn.com/magazines/fortune/global500/2007/`, *retrieved 2008-06-09*

[16] `http://tidy.sourceforge.net`, *retrieved 2008-07-27*

use of JavaScript and Flash, thus these were excluded from the further evaluation, resulting in a corpus size of 44 web pages.

4.5.2 Evaluation Design

Five different error classes were identified that can occur on automatic segmentation in the given scenario:

Missing segments are the parts of a page that should be marked as an own segment but are not recognized by the algorithm.

Superfluous segments are segments that are found but are not segments according to the definition given in section 4.3.1. For example, this could be a nested segment that adds no textual information.

Incomplete segments are segments that are principally correct but there is a part of the segment not included. An example is a textual part of the page that lacks a heading.

Too big segments are segments that are too big in respect to the segment definition, e.g. a segment spanning two comments in a blog.

Wrong segments are segments that are completely wrong and do not match any of the criteria above.

Evaluating a segmentation algorithm needs *human judgements* as a reference. Thus, five participants (between the ages of 23 and 28, all male including three students of Information Science, one student of Education and one member of research staff) were introduced to the definition of coherent segments. The participants were instructed to examine each processed web page and check whether the found segments are correct or can be classified as one of the above–mentioned errors. The participants were given a short introduction to HYRECA and shortly trained to interpret the coloured output. On inspecting the segmented web pages, they were allowed to switch between the three visualizations, the original web page and the two versions highlighting the segments found by HYRECA. Finally, the participants were asked to count the occurrences of each of these errors. On average, an evaluation took about three hours for each participant.

The rating values (i.e. errors found per segmented web page) were converted into the ratio of error r_e in relation to all segments found on the web page for each web page, rater and error class. The ratios were aggregated in five ordinal classes denoting the severity of error in fine–granular steps of 2.5%, i.e. *category 1* as $0.0 <= r_e <= 0.025$ (“completely correct segmentation to 2.5% of all segments were wrong”), *category 2* as $0.025 < r_e <= 0.05$, *category 3* as $0.05 < r_e <= 0.075$, *category 4* as $0.075 < r_e <= 0.1$ and *category 5* as $r_e > 0.1$. This categorization is based on the observation that the participants found an error rate of more than 10% (i.e. 10% of all found segments were erroneous, which is category 5) not to be acceptable anymore.

4.5.3 Results of the Evaluation

The data gathered in the evaluation has been evaluated with regard to the participants’ *agreement* on their ratings and the *quality* of the segmentation approach.

Agreement between Raters

With the five categories defined above, the agreement p that denotes the agreement in percent of the cases can be calculated (see equation 4.1), with n being the number of raters, k the number of categories indexed by $j = 1, \ldots k$, and n_j being the number of ratings assigned by all raters to the jth category.

$$p = \frac{\sum_{j=1}^{z} n_j^2 - n_j}{n(n-1)} \tag{4.1}$$

The agreements of the participants of the evaluation can thus be measured for each error class per web page. A weighted mean agreement of $\bar{p} = 0.70$ (with standard deviation of 0.05) resulted for all of the 44 web pages taken into account in this evaluation.

Genre	$\bar{p}$	Standard deviation
Blogs	0.74	0.20
Companies	0.69	0.26
Shops	0.60	0.19
News	0.72	0.23
Miscellaneous	0.73	0.17
Weighted total	0.70	0.05

Table 4.1: The mean agreements $\bar{p}$ per genre

This value shows that, albeit the raters agree in many ratings, that the task of rating coherent segments is subjective. Especially when incorrect segments were identified, there were different perceptions how the problematic segments should have been partitioned. Another source of disagreement were the error classes, e.g. the error classes *superfluous* and *completely wrong* were often confounded.

A per–genre value of agreement is given in table 4.1. Here, especially the genre *shops* proved to be difficult to rate, probably due to the participants' uncertainty how product presentations were to be segmented.

Correct Retrieval of Segments

Table 4.2 shows the different types of counted errors, averaged over all pages. The low number of segments marked as too big and the high number of superfluous segments indicate that HYRECA is tending to produce segments that are too fine–grained. Simultaneously, the high number of missing segments shows that the approach presented here is improvable, but just adjusting the thresholds that are used to differentiate valid segments from too fine–grained fragments will not work, as this would in turn increase the number of superfluous segments.

Precision and *recall* are metrics for measuring the quality of IR and extraction algorithms that are commonly used. Precision is the proportion of retrieved items that are actually relevant, recall is the proportion of relevant items that are retrieved [8]. Applied to page segmentation, precision is the proportion of detected segments that are actually applicable to the segment definition and recall is the proportion of segments that are correctly detected. In order to calculate precision and recall, the results that are based on single web pages are averaged per genre. The following notations are used:

Type of error	Mean	Standard deviation
Superfluous segments	1.66	2.52
Missing segments	1.07	1.92
Incomplete segments	0.93	2.34
Too big segments	0.20	0.77
Completely wrong segments	0.24	1.92

Table 4.2: Average errors per web resource reported by participants of the study, listed by error class.

n_d The total number of segments in a web page detected by HYRECA.
n_a The total number of segments in a web page as given by the participants. This number is estimated based on the ratings of the participants.
n_c The correctly identified segments of a web page.
e_m The average number of *missing* segments.
e_s The average number of *superfluous* segments.
e_i The average number of *incomplete* segments.
e_b The average number of segments that are *too big*.
e_w The average number of segments that are *completely wrong*.

The actual amount of segments n_a in a web page is an average of the participants' ratings, because the exact number of segments is subjective and there is no absolute gold standard. Thus, n_a is approximated by the number of detected segments plus the average number of missing segments with the average number of superfluous segments subtracted:

$$n_a = n_d + e_m - e_s \tag{4.2}$$

Further, the number of correctly detected segments n_c is the number of detected segments without all erroneous segments:

$$n_c = n_d - (e_s + e_i + e_b + e_w) \tag{4.3}$$

With equations 4.2 and 4.3, precision and recall can be defined as

$$precision = \frac{n_c}{n_d} \qquad recall = \frac{n_c}{n_a} \tag{4.4}$$

Using equation 4.4, the quality of HYRECA can be calculated per web page and thus be averaged to an appraisal of the quality in the different genres (see table 4.3). The weighted average of precision is 0.86, recall is 0.88. These results are feasible when considering the fuzziness of the manual segmentation task. Notably, the *miscellaneous* category does perform below average. This is because this category is a collection of assembled web pages, some of which were selected because they pose challenges to a segmentation and some that do not expose a very refined structure, e.g. wiki pages that largely consist of plain paragraphs and therefore provide only few pattern structures.

Genre	Pages	Precision	Recall
Blogs	10	0.95	0.96
Companies	7	0.80	0.84
Shops	8	0.93	0.95
News	9	0.93	0.94
Miscellaneous	10	0.70	0.73
Weighted total	44	0.86	0.88

Table 4.3: Precision and recall of correct segmentation

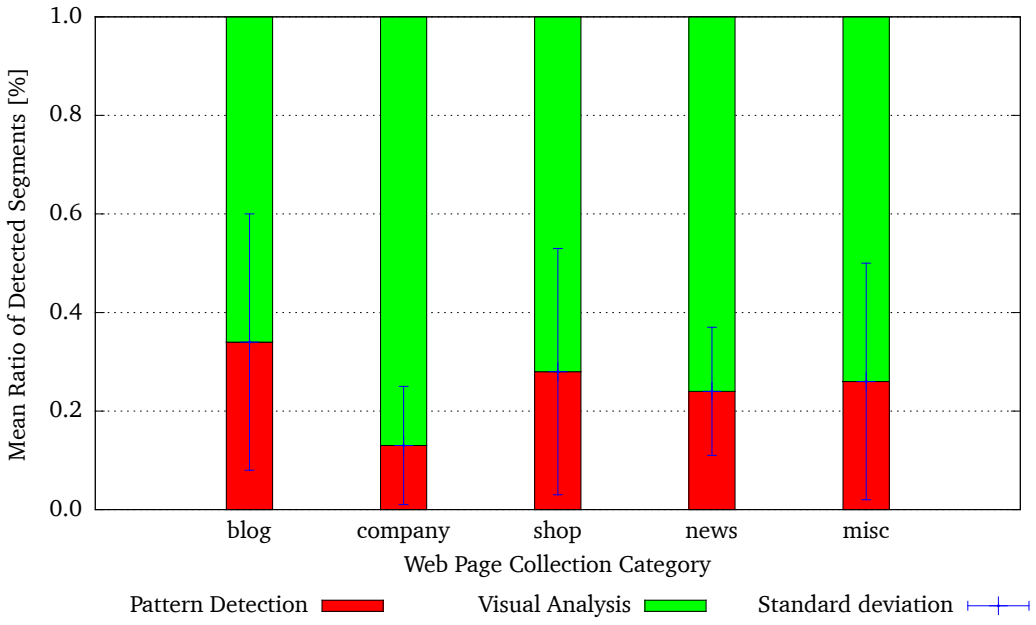


Figure 4.12: Average ratio of segments found by the pattern detection vs. the visual approach per genre. Note: The ID and class heuristics have been omitted in this diagram, as they yield constant results per genre.

Further, the ratio of segments recognized by the pattern detection vs. segments detected by the visual analysis are compared in figure 4.12 for each genre. The standard deviation shows that the ratios are affected by large deviations between the web pages in each genre. Therefore, applying both approaches seems to be a good strategy to provide a good detection of segments in a diverse range of web pages. A limitation of this evaluation is that it does not consider the order of the applied detection steps, e.g. the ratios in the figure would probably different if the visual approach would be applied first. This, however, would have necessitated the participants to rate the segmentation of each web page twice, but this could not be expected of the participants due to expenditure of time.

4.6 Conclusions and Outlook

In this chapter, the novel approach to web page segmentation HYRECA has been presented. Related work in the field of automatic web resource segmentation has been analysed and the shortcomings of the respective approaches have been shown. Based on the analysis of related work and its deficits, design goals for a novel approach were derived and the HYRECA algorithm was presented that combines a pattern detection approach, a visual approach and a class/id heuristic approach. An evaluation methodology was proposed that adequately captures the definition of coherent segments and results of a user study were presented. The results show that, depending on the genre of a web resource, good results can be achieved.

Although targeted at providing usability and retrieval support in ELWMS.KOM, HYRECA's value as a pre–processing step in different application scenarios is considerable, e.g. in filtering irrelevant content like advertisements, small–screen display of web resources and retrieval strategies targeted at relevant content. Further, the notion of *patterns* that represent reoccurring structures in a web resource's DOM is a feature that is applied in the object of investigation in chapter 5, where patterns are used as hints for automatically determining the structural genre of a web resource.

5 Web Genres as Metadata

As chapter 2 has shown, tagging is an adequate and light–weight method of assigning metadata to LRs. Many researchers have stated the value of automatically extracting metadata from content [85, 144], mostly focusing on topical metadata [143, 131, 180]. However, another metadata class that is relevant to users is the *type* of a resource. This is reflected in ELWMS.KOM by providing the *Type* tag that encompasses metadata describing the genre or physical attributes of a resource, e.g. whether the resource represents a *PDF* file or if it contains an *blog* post. Often, this information is important for users, and as it is orthogonal to a resource's topic and provides additional hints about the form of a resource, it can benefit the retrieval of resources [26]. In many cases, this classification is trivial to be automated (e.g. for the file format of a resource, cf. [22, 31]), but there are types that are often used (specifically the *web genre* of a resource, cf. section 5.1.1) but not easily distinguishable for a classification approach.

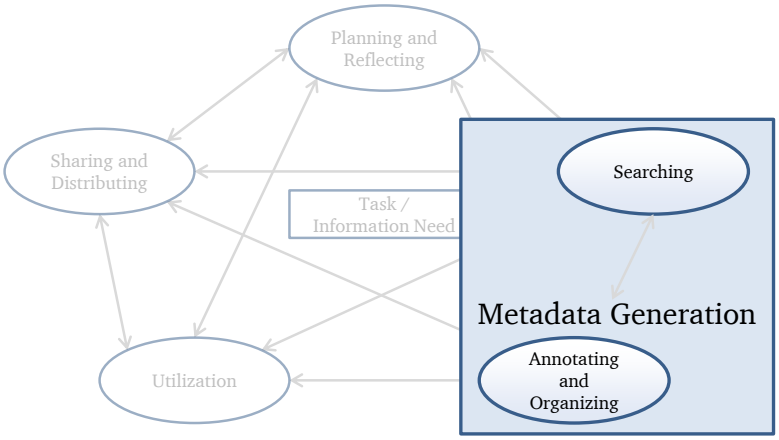


Figure 5.1: Supporting Resource–Based Learning by providing metadata benefits the Retrieval and Organization processes

In RBL, an approach to automatically derive metadata from a LR can be classified as a means of annotating and structuring the found resources (see figure 5.1). It helps learners to create a consistent vocabulary in their resource organization and therefore facilitates both the structuring and the retrieval process. In addition, a consistent vocabulary helps other learners to discover resources efficiently. A common example for an inconsistent vocabulary is in social bookmarking applications that weblogs are often tagged with the different terms `blog`, `blogs`, `weblog` and `weblogs` (cf. section 5.1.1). An automatic approach normalizes this different use of terminology and therefore unifies the tagging vocabulary. This automatic approach is applied when a learner saves a web resource to his knowledge network in ELWMS.KOM (for example, see figure 5.2 where a web resource of the web genre *Wiki* has been recognized).

This chapter introduces *web genres* as an assignment type of Learning Resources. It examines the applicability of the different employed features, introduces a novel feature types, e.g. representing the structure of a web resource, and presents and evaluates a language–agnostic approach called LIGD to detect one of the targeted web genres *blog*, *wiki* and *forum* of a web resource.

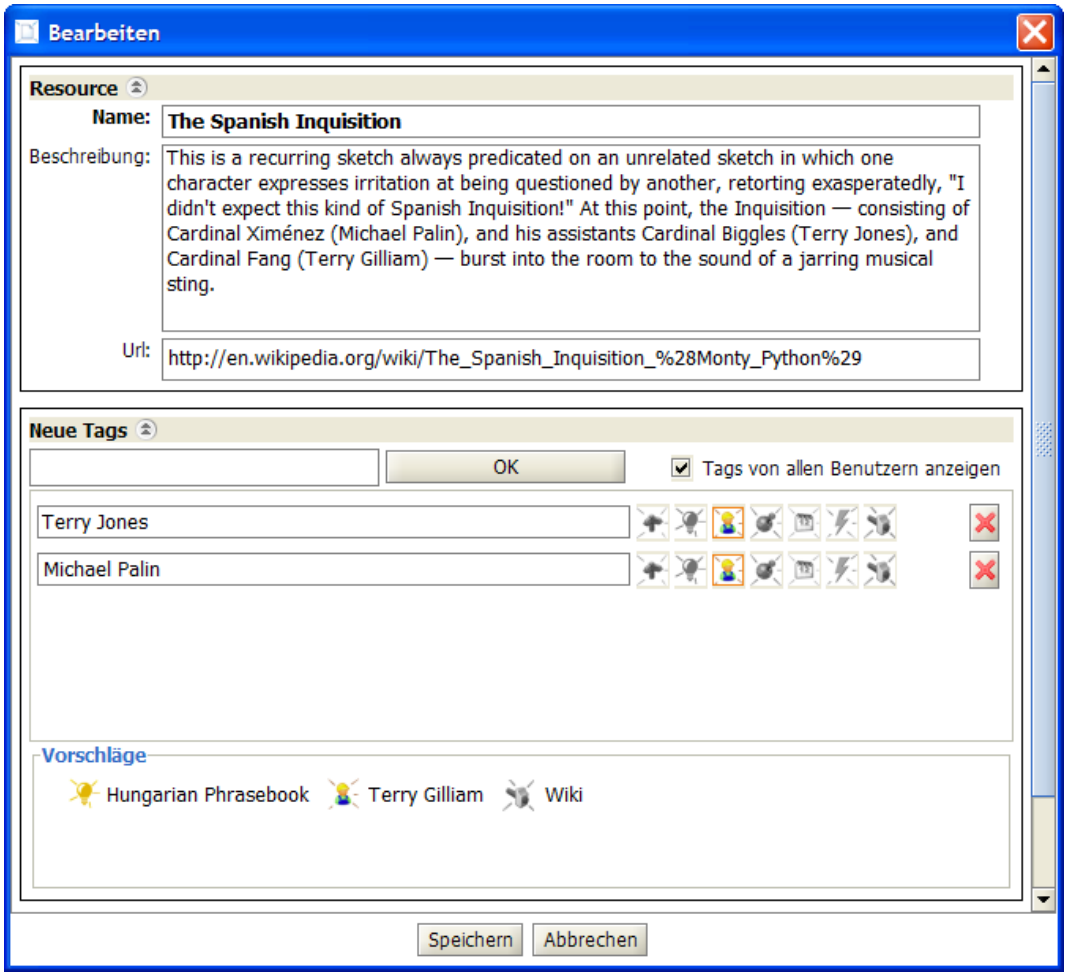


Figure 5.2: A web resource is saved, and ELWMS.KOM recommends the tag “Wiki” with the type *Type* (see panel “Vorschläge”).

5.1 Introduction and Motivation

In ELWMS.KOM, learners are collecting web resources that meet their specific information needs. However, not all web resources are homogeneous: different applications have different paradigms and conventions of usage, content creation, authorship, annotation (e.g. commenting) and reception, and therefore provide different kinds of information in different ways. Web resources can be grouped into self–similar classes (subsequently called *web genres*) based on these properties. For example, web resources have diverse functionalities (and thus can be assigned different genres): displaying a product (e.g. the web genre *e–shop*), representing a person in a social or organizational context (e.g. *personal* or *academic homepage*) or allowing to follow and take part in discussions in a *forum*.

In ELWMS.KOM, the *Type* tag type has been introduced in order to capture this diversity of different web genres. As, section 3.1.2 shows, in ELWMS.KOM, the multilingual nature of LRs is an additional challenge. Thus, an approach that automatically determines the web genre of a web resource should be able to handle resources in different languages and therefore do not depend on linguistic features of a resource.

5.1.1 Examination of Tags denoting Web Genres in Social Bookmarking

The differentiation between different text and web genres is valuable for personal organization of knowledge. For example, genre information is commonly utilized in *tagging* applications, and Golder and Huberman [81] name *genre tags* among several different other utilization possibilities of tags:

Topic tags represent the overwhelming majority of tags used in tagging systems, as they identify what (or who) a resource is about.

Genre tags identify the kind of resource in addition to the topic. For example, web resources can be tagged as `example`, `article` or as a web genre like `blog` or `wiki`.

Content ownership tags represent the entity that owns or created the tagged resource, e.g. `W3C` for a web resource containing the W3C HTML 4.0 standard.

Tags identifying qualities or characteristics often denote subjective opinions about the tagged resource, e.g. `cool` or `funny`.

Refining tags do not stand alone but refine other tags, often numbers or years are used, e.g. `2010`.

Self–reference tags are used to identify content in terms of its relation to the tagger, e.g. `mystuff`.

Task organization tags serve to group resources needed in a related task and structure the task of a tagger, e.g. `toread` or `jobsearch`.

Respectively tags that denote topics of resources and tags that encode the genre of resources are often used. The latter contain — to a high degree — different text and web genres. This can be observed in different tagging applications, especially in social bookmarking applications. For example, in a subset of bookmarks of Delicious (crawled in March 2008, containing approximately 22 million bookmarks tagged with 64 million tags of over 41,000 users — that is about 527 resources per user and 1.2 million unique tags), the web genres *blog*, *wiki* and *forum* occur in the top 200 tags. Especially the tag `blog` (including its synonyms `blogs`, `weblog` and `weblogs`) makes up 14.01‰ of all tags, making it the tag with the highest frequency. Also, for example, the web genres *wiki* (1.90‰) and *forum* (0.98‰) are frequently used. Further, text genres are equally important, e.g. the tags `reference` (8.21‰) and `howto` (5.77‰) are the most important for denoting text genre. Thus, a substantial part of tags used for tagging web resources represents text or web genres. The 50 tags with the highest frequency are listed in table C.3 in

appendix C. For comparison, figure 5.3 shows that the most current popular tags used on Delicious have not changed substantially since 2008, only their ranking order has changed slightly.

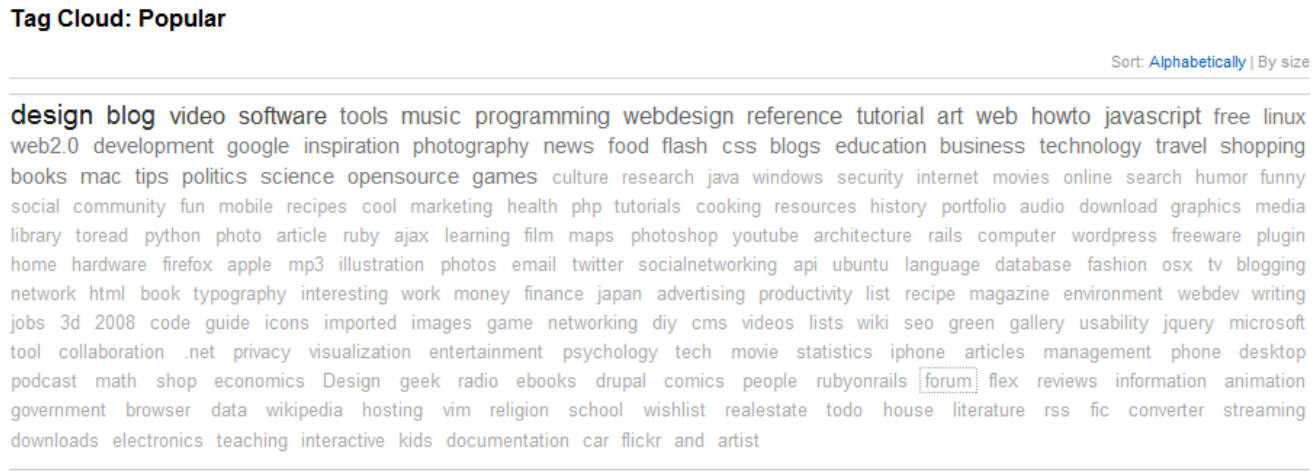


Figure 5.3: Tag cloud of the most often used tags of Delicious as of 2011-01-17 according to `http://www.delicious.com/popular/`

However, users often tag inconsistently [124, 81, 36]. For example, this can be observed with different tags (e.g. `blog`, `blogs`, `weblog` and `weblogs`) denoting the same web genre. Thus, automatically detecting genre tags for web resources is helpful, as it standardizes the tag names and enriches the metadata used for describing web resources.

For supporting the user to attach consistent tags to web resources, many approaches have been presented that automatically extract topical tags from web resources and recommend these to the user (e.g. [39, 52, 133, 180]), often taking into account personal tagging behaviour or an existing folksonomy. However, detecting the web genre of a web resource in contrast to topical tags is not possible based on content words or already assigned tags exclusively. Thus, a different approach has to be followed.

5.1.2 Other Scenarios for Web Genre Detection

Besides tagging, typical use cases for automatic detection of web genres can be found in all fields where analysing huge amounts of unstructured information from the web is involved: for example, in the field of IR, users searching for information on the web may get more specific search results, if they are able to state what kind of information in what type of page they expect as a result [60]. Further, web pages contain more than the pure information to be extracted, e.g. usually there are navigation elements, headers or footers. Because this adds so–called noise to the information space that is analysed and searched [201], knowing the genre of a web resource enables to apply specifically tailored pre–processing steps, thus reducing noise effectively.

Another usage scenario for web genre metadata is *Community Mining*, i.e. analysing network structures of communities of users in social software applications and extracting properties and the structure thereof [43]. A popular approach to Community Mining is crawling the web pages of one or multiple social software applications (e.g. a wiki or a subset of the blogosphere) by following hyperlinks to other web pages and extracting the desired information. Here it is important to know the genre of the linked web resources in order to use heuristics tailored to the according web genre. For example, a typical wiki

article has several authors that edit the same content, whereas a blog post normally has only one single author, thus authorship of content has to be identified in different ways.

5.1.3 Structure of this Chapter

In this chapter, the challenge of recognizing and distinguishing the web genres *wikis*, *forums* and *blogs* is targeted, as these are content–creation backbones of online communities, widely adopted and used, and support different paradigms of content creation and collaboration. Further, as they often contain informative content, web resources of these genres are often bookmarked for later reference, especially in self–directed learning settings. The contributions of this chapter are as follows: section 5.2 presents an overview of other approaches to recognize the web genre of a given page. The pattern features for each web genre to classify are derived in section 5.3 and novel pattern features are proposed that do not depend on a linguistic analysis of the web page's content but rather analyse the structure of the web resource's HTML markup. Thus, this novel approach called LIGD is fully language agnostic and independent of specific language analysis tools (e.g. part–of–speech taggers). Further, section 5.4 presents a corpus that reflects the choice of web genres, including resources in different languages and provided by different applications. In section 5.5 an evaluation of LIGD is given that shows that this approach performs well with 94.3% sample pages correctly classified. Finally, section 5.6 gives a conclusion and presents perspectives for future work.

5.2 Related Work

Genre is a term widely used in rhetoric, literary theory, media theory and linguistics to refer to a distinctive type of work (particularly texts) [50]. *Genre Theory* provides a framework for classifying and grouping these works into well–defined taxonomy schemes.

As Kessler et al. [101] describe the term *genre* in the context of texts, it "is necessarily a heterogeneous classificatory principle, which is based among other things on the way a text was created, the way it is distributed, the register of language it uses, and the kind of audience it is addressed to".

Building on Kessler's terminology, the term *web genre* is used to describe and classify web resources by structural, functional, contextual and institutional characteristics like style, form, content and use of language. It spans use, intention and display of web resources and is considered orthogonal to the *topic* of a web resource [27, 26].

In contrast to web genre classification, *web page classification* ([190, 163]) takes into account a topical classification taxonomy scheme. For example, web page classification assigns web resources to different topics like "Family" or "Garden" [199], whereas web genre classification assigns them to functional classes like "e–Shop" or "personal homepage" [132].

5.2.1 Ambiguity of Taxonomies and Evolution of Web Genres

Chandler [50] notes that classifying genres into a hierarchical genre taxonomy is not a neutral and objective procedure. This means that there are no "right" or "wrong" genre taxonomies, their use depends on the point of view of the researcher.

Depending on the anticipated use case, coarse or fine–grained categories for genres are developed (e.g. *blog / wiki* versus *personal / academic homepage*) that base on different properties of a web resource. However, even for human beings, recognizing and classifying web genres is not easy and heavily depends

on unambiguity of said categories and planned use [132]. For example, how would a *bliki* (a hybrid genre form of blogs and wikis) be categorized if only blogs or wikis were available as classes?

Further, although genres follow certain conventions and quasi–standards internally, they may develop aspects that are different between web resources of the same genre [27, 168]. For example, a particularly influential factor is time: genres evolve, their presentational and structural properties follow a possibly transformed use case and new trends appear (especially regarding technology). This affects web genres in particular: for example, blogs first were used as a platform to publish informal articles to a small circle of friends and family, similar to a personal diary. With the advancement of the Web 2.0 hype, content publishers like newspapers and corporations started to use the medium blog as the means to easily distribute their contents, the blog entries being written in more formal style of language. Thus, the style of writing has changed considerably over time, making it difficult for automatic text genre detection to derive the correct genre. Further, complex web pages used to be based on table layout. Nowadays, however, the paradigm of semantic HTML gains more and more importance and acceptance of new technologies, e.g. support of CSS2 in major web browsers, enables authors of web resources to layout web pages by other means than tables, affecting the presentational properties of pages with the same content now and then considerably. Eventually, since it has become possible to embed content from other third parties like comments[1], the structure of the web genre blog has changed substantially, as it is not necessary to provide an own comment system anymore. This lack of a comment system changes the layout of the blog page. Thus, all approaches to detect the genre of a web resource based on its structure are only valid while the underlying genre itself does not evolve too significantly.

5.2.2 Related Approaches

Related approaches differ in *web genres* taken into account, *features* (also named *cues* by some researchers) used for classifying the web genres and *machine learning algorithms* they use for the actual classification task.

Detecting a genre of electronic texts has been a focus of research since the late 1980s (cf. [101]). These approaches mainly focused on plain text, thus they primarily applied linguistic analysis and some structural metrics (punctuation, sentence–length, readability metrics like the Flesch metric [73]) in order to identify the genre [60, 71, 72]. In contrast to *text genre* identification, *web genre* identification additionally may take into account structural features provided by the markup of HTML–based web resources.

In a user study, Meyer zu Eissen et al. [132] identify seven web genres that are relevant for users' expectations towards the genres of search results when looking for information using search engines. They conclude with the following web genres fulfilling the users' expectations: *Help* (e.g. Frequently Asked Questions (FAQs)), *Article* (e.g. scientific articles, but they also subsume all longer web pages with continuous text), *Discussion* (forums, mailing lists), *E–shop* (product presentations and sale), *Non–private Portrayal* (presentation of enterprises and public institutions), *Private Portrayal* (personal homepages), *Link Collections* (documents that largely consist of links) and *Download Pages* (pages that offer software to be downloaded). The authors consider use cases that require classification in real–time (online), so they state that the features used for classifying must not be too complex or computationally expensive to collect. They differentiate between three groups of features: document terms (e.g. frequencies of words, number of spelling errors, *closed class word sets*, i.e. typical keywords for each genre), linguistic features like part–of–speech (POS) analysis and syntactical analysis and simple text statistics (e.g. frequencies

[1] e.g. Disqus (`http://disqus.com/`, *retrieved 2011-02-07*) provides external hosting of comments

of punctuation). Based on these features they develop two feature sets *A* and *B* that can be primarily discerned by computational costs: feature set *A* is based on analysis of certain markup elements, simple text statistics and genre specific closed class word sets. Feature set *B* is based on POS and linguistic analysis. Multi–Layer Perceptron neural networks and Support Vector Machines (SVMs) are used to actually classify the genres and the results range between 60% and 80% accuracy with feature set *B* slightly outperforming the computationally inexpensive feature set *A*.

Dewdney et al. [60] focus on the web genres *Advertising*, *Bulletin Board*, *FAQ*, *Message Board*, *Radio News*, *Television News* and *Reuter Newswire*. They compare the use of presentation features, word features, as well as the combination of using Naive Bayes, C4.5 decision trees and SVM classifiers. Presentational features reflect the way information is presented. Relevant word features are individual words weighted by tf–idf (term frequency — inverse document frequency), a statistical measure used to evaluate how important a word is for a document in relation to its occurrence in a collection or corpus. Then the word features are filtered by estimated Information Gain (a method to reduce computational complexity in machine learning by reducing irrelevant features). Here, Dewdney et al. subsume linguistic features (e.g. use of tense, prevalence of adjectives as detected by POS tagging), layout features (e.g. line–spacing and non–alphanumeric characters) and miscellaneous features (e.g. readability measures, average length and sentence length and punctuation). For the seven mentioned genres they experimentally gain 92% classification accuracy.

Amitay et al. [2] not only classify genres of web resources but also genres of whole sites, like *Enterprise Web Sites*, *Media Sites* (e.g. sites of major TV stations and newspapers), *E–shops* and *Universities*. They use structural features that result from site and link structure on several web pages of the site only. Ingoing links, internal links and outgoing links are put into relation to the same page, the same site and external resources on the web. The performance of this approach is satisfactory with about 60% correct classifications. This allows to draw the conclusion that classification of genres of whole websites is possible by only taking into account the link structure between pages.

A comprehensive compilation of web genres is found in the works of Santini [167, 168, 169]. She presents a selection of seven representative web genres: *blog*, *e–shop*, *FAQs*, *online front–page* (e.g. main pages of universities, enterprises and public organizations), *listings*, *personal homepage* and *search engine pages*. Santini discerns — along the lines of Meyer zu Eissen and Stein [132] — three different sets of features that she uses for classification: linguistic facets (similar to part–of–speech word sets), word frequencies (based on bag–of–word sets) and likewise structural information in HTML markup. The features deemed as most effective are closed class word sets and markup information (e.g. size of page in characters, number and frequencies of HTML tags and internal / external navigability through hyperlinks). Besides these, other features like POS trigram frequencies, HTML facets and several linguistic facets are evaluated. A concise summary is available online at the author's web page[2]. In [169], Santini proposes a multi–classification approach, as she states that often a single genre is not enough when there are ambiguous genre classes. Additionally, as genres are evolving, she postulates to follow a more flexible, dynamic approach of genre classification, allowing the assignment of none, one or multiple genres to a web resource.

Levering et al. [118] investigate whether visual features of HTML web pages can improve the classification of fine-grained genres instead of focusing on text–based features. They purposely choose their genres (*store homepages*, *store product lists* and *store product descriptions*) to be easily distinguishable in order to focus on feature construction. Besides textual features (readability measures, bag–of–words, POS and text statistics) and HTML features (link, form and HTML tag counts, scripts and URL features)

[2] `http://www.itri.brighton.ac.uk/~Marina.Santini/`, *retrieved 2008-10-21*, see section "Genre Features"

they heavily apply visual features like image counts and statistics (e.g. average of all image sizes), area statistics (e.g. total areas of different object types, their relative percentages in comparison to the total area of the object type) and placement statistics (e.g. location distribution of terms). Their use of all these (over 12,000) features results in classification accuracies between 85% and 95% using SVM.

5.2.3 Approaches to Classifying Blogs

For the approach presented in this thesis, LIGD, the following publications are relevant as they focus on classification of a single genre (here: *blogs*) and have a similar understanding of web genres as presented in this chapter.

Nanno et al. [142] focus on recognizing Japanese blogs and blog–alikes. Their goal is to identify blogs, i.e. classifying into *blog* or *not a blog*. Their approach is based on the observation that one of the most significant properties of a blog is that blog systems usually present their posts in temporal linearity, i.e. that dates and times that are presented are either in strictly ascending or descending order. Thus, consistently formatted date strings serve as features for recognizing blogs as well as delimiters for detecting single blog posts. Instead of machine learning, Nanno et al. use manually created patterns for the identification of these date strings and genre determination, claiming accurate classification in 84% of all cases.

Elgersma et al. [65] focus on markup based features (e.g. number of HTML comments) and closed class word sets like “archive” or “comment” in order to recognize blog entries on arbitrary web pages. Classification is binary, i.e. only the fact whether a given page is a blog or not is recognized. Furthermore their algorithm tests given pages on occurrence of links to some of the 20 most–used blog hosting providers (which is arguably not a very stable feature, as providers come and go). On these rather simple features different techniques and machine–learning algorithms are executed. According to the results, Elgersma et al. propose using Support–Vector–based methods for comparable classification tasks (e.g. SVM, with approximately 93% accuracy), although there is no significant difference between most of the other 17 learning algorithms which have been applied (between 88% and 93% accuracy).

Kolari et al. [106] mention that blogs as a publishing mechanism have crossed international boundaries and therefore blog posts are often written in another language than English. They even report blogs that are multilingual, i.e. blog posts of the same blog are written in different languages. Therefore they introduce an additional feature called *bag–of–n-grams* besides *bag–of–words* that converts text into tokens of characters with a certain window length. This enables their approach to take into account the often similar word stems in different related languages (e.g. the English `comment` and the German `Kommentar` still have the four–gram `ment` in common and thus may provide further cues on classification). However, they do not mention how good this approach is with respect to detecting the genre of multilingual web resources.

5.3 Features Used in Language–Independent Web Genre Detection

Nearly all of the approaches presented in section 5.2 at least partially rely on linguistic analysis of the web document’s content. This is feasible, as their use cases cover web genres that use language in different ways. For example, *advertisements* use a different wording than *personal home pages*.

However, the web genres *wikis*, *blogs* and *forums* which are focused in LIGD are types of web applications that are wide–spread and utilized internationally, and thus contain information written in different languages. Therefore, use of language and linguistic features are not necessarily discerning features and should be neglected in favour of a truly language agnostic approach.

The following sections describe the feature classes that are used in this novel approach to web genre classification in detail.

5.3.1 Pattern Features

The focused web genres raise certain expectations concerning presentation of information. Even if humans are not able to understand the language of a web document at all, they are — to a certain degree — still able to visually differentiate between these genres and classify web pages accordingly (see figure 5.4).

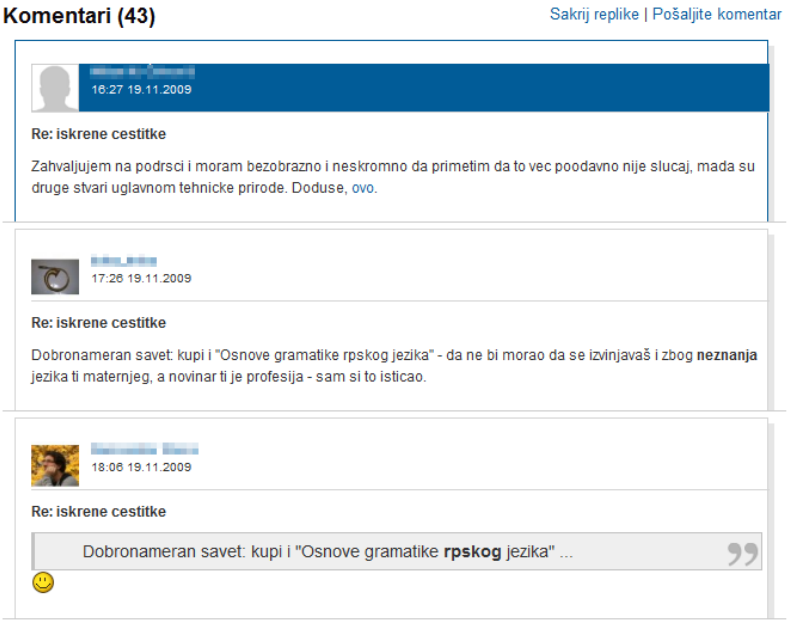


Figure 5.4: Example of a blog's comments following a common structure, thus being identified as patterns (marked by dotted line). Although the text itself is in Czech, it is easily recognizable by the structure alone that this picture shows some kind of comments.

This is due to the typical content structure of a web resource of a specific genre, e.g. a blog basically consists of blog posts and a number of comments to each of this blog posts. This can be easily perceived by a human, as this structure is mirrored in the visual layout of the content blocks [189, 118]. The same content blocks are rendered into HTML markup in the same way by the blog's template engine, meaning that all comments on a blog post page share a common structure, only the user generated content (i.e. the text entered by a user) contained in this structure varies (e.g. multiple paragraphs or additional links or images). Thus, in the best case (for example, if a blog post has two or more comments), similar markup is repeated. This repeated markup structure is called a *structural pattern*. The pattern extraction method is based on Rafiei et al. [153], a computationally affordable approach to measure (sub)–tree similarity. Rafiei et al. use it to describe a concise and accurate structural summary of XML documents. Using this description allows to detect documents sharing a similar structure efficiently. These similar structures are then used as pattern candidates. This notion of *pattern* is equivalent to the pattern definition presented in chapter 4.4.2.

Based on these structural patterns, some novel features that serve to identify selected web genres are proposed: these features represent properties of a web resource that relate to the number, size,

hierarchy, structure, in–page location and content ratio of identified patterns. The following list presents an overview of the different pattern feature types.

Number of patterns The feature `pattern_nr` represents the number of detected patterns in a web resource. The feature `pattern_outer` indicates the number of patterns that are not enclosed in other patterns and thus hints on the hierarchical structure of patterns. If there are no hierarchical patterns, `pattern_outer` equals `pattern_nr`.

Size of patterns This feature group encompasses the features `pattern_size` and `pattern_median` which indicate the mean and median size of the patterns in bytes. The median feature is especially interesting, as it represents the size distribution of patterns.

Location of patterns The only feature in this group is `pattern_start` that denotes the ratio of text in a web document before the start of the first pattern occurrence. It is represented as a value $p_{\text{start}} \in [0..1]$.

Content ratio of patterns This feature group represents the fraction of a page that is contained in patterns. In order to calculate it, all content of outer patterns (i.e. patterns that are not contained in other patterns) is aggregated and contrasted to the overall web resource content. There are two different types of content ratio patterns: `pattern_ratio` denotes the fraction of all content (both HTML markup and text) in comparison with the whole web resource, whereas `pattern_ratio_text` only takes into account the textual content of the web resource. Both features give information about the patterns' coverage of a web resource and are represented as a value $p_{\text{ratio}} \in [0..1]$.

Pattern Hierarchy This feature group is especially targeted at web genres that make strong use of nested patterns by building a hierarchical structure of patterns. For example, blog pages often contain nested comments, which would be represented by this feature. `pattern_depth_mean` denotes the average pattern hierarchy height and `pattern_depth_median` represent the median of pattern hierarchy.

5.3.2 Tag Frequency Features

Tag frequencies are the most basic features that are utilized in most of the related work (e.g. [2, 132, 168]). They are mere counts of the occurrence of HTML tags. Here, all tags defined by the W3C's HTML v.4.01 specification [154] are used (cf. table B.1 in appendix B). Deprecated tags (e.g. `blink`) and HTML v.5 [88] tags (e.g. `section`) are ignored.

Based on the tags contained in the web resource, additional features have been introduced in this work:

Class frequencies denote how many HTML tags have a `class` attribute attached to them. Classes serve to assign a style to a tag so that all tags having this style can be visually manipulated by CSS or easily accessed by ECMAScript. This is often used in blogs and forums (e.g. for styling recurring blocks of contents like comments) and thus this feature serves to differentiate between these genres and wikis.

ID frequencies is the number of unique IDs that have been assigned to tags. These IDs serve to access tags that are unique, i.e. only one ID may exist in one web resource. They are commonly used to denote larger blocks of content, e.g. for identifying a web resource's header or body.

The block tag ratio represents the fraction of tags that are *block level elements* (cf. chapter 4.2). This feature serves to differentiate between web resources that have many block level elements and therefore exhibit a distinctive structure.

5.3.3 Facet Features

HTML Facet features [167] are derived from the tag frequencies. They aggregate tag counts into functional groups, e.g. accumulating frequency counts of tags that define layout properties of the web resource. There are three facets that build on tag frequencies:

The layout facet groups tags that relate to the layout and logical formatting of text.
The typographic facet aggregates all tags that affect the typography of the text.
The functionality facet is mainly related to the possibility of user interaction with web resources.

For a listing of the respective HTML tags that are aggregated in each of the facets, see table C.1.

5.3.4 Link Features

Further, Santini [167] describes three facets that capture information about the navigability of a web resource and are special cases of the feature that captures link frequency (i.e. occurrences of the `a` tag):

The general navigability facet contains the count of links that provide navigation to another web page or to anchors of the same web resource. The definition of anchors (i.e. by using the `a`–tag with the parameter `name`) is excluded.
The external navigability facet contains only links to external web resources, e.g. to other `http` or `ftp` pages.
The internal navigability facet contains the number of links to locations that are on the same web resource, i.e. internal links.

In LIGD, some additional, novel link features have been developed. Based on the assumption that pure counts of tags do vary too much between web resources of different lengths, link ratios are introduced into the feature set that set the number of respective link classes into relation with the total link count of a web resource. These features are designed to add to the respective navigability facets.

The anchor link ratio denotes the ratio of `a`–tags that only define an anchor. This feature serves as a hint on the structure of a web resource, for example a wiki page often contains multiple sections that are navigable via a table of contents. Thus, the anchor link ratio for this genre could be higher than e.g. a forum start page.
The page link ratio represents the ratio of `a`–tags that link to the same web resource (internal links). This feature is similar to the anchor link ratio, but it represents the link's source, not its target.
The site link ratio represents the ratio of links to the same web site. It captures on–site navigation and is assumed to be high for all genres targeted in this work. Its primary use is to discriminate between the targeted web genres and arbitrary web pages.
The external link ratio denotes the ratio of links to external, i.e. off–site targets. Its function is to separate the different sub–genres, as e.g. a blog start page will link to its own content, whereas a blog post page will have multiple comments that link back to their author's web sites.

5.3.5 Content Features

These features take into account the textual content of the web resource. While the linguistic features like use of language or certain keywords are ignored, these features exploit the *structure* of the language itself. There are three different features in this category:

The number of words is an indicator of the textual length of the web resource and has been previously used by other approaches [132, 167].

Punctuation frequencies are important features in text genre classification [60]. They consist of a simple count of occurrences of punctuation symbols that are internationally used (i.e. that are contained in the American Standard Code for Information Interchange (ASCII)[3] character set). Examples for this punctuation are `.`, `!` and `(`.

The text/markup ratio is a feature that represents the balance of text that is displayed in the browser and the size of HTML markup. It is based on a similar feature described by Rehm [158]. It is an indicator of the weight of markup compared to the actual text and is assumed to be lower when a lot of markup structure is present. For example, forum thread pages have usually a low text/markup ratio because there is a lot of `table` boilerplate code, whereas wikis are commonly light–weight and thus have a high text/markup ratio.

5.3.6 URL based Features

A web resource's URL often exposes information about the web genre of that page. This feature group reflects this information.

URL path length In related work especially the URL length in characters is used [118]. However, in this work, this feature is altered to represent the URL path length, e.g. the path length of `http://example.com/2011/01/20/title-of-the-post.html` would be three, as the web resource has three parent "folders". The reasoning behind this feature is that especially blogs often use the URL paths to build a hierarchy based on dates, whereas wikis usually have a flat path hierarchy.

URL date containment is a feature that represents whether a date fragment could be detected in the path. For example, in `http://example.com/2011/01/20/title-of-the-post.html` the year string "2011" can be found and a string that may represent the first month of the year. This feature is dependent on a cultural bias, as the year 2011 can be represented as 5771 in the Jewish era. However, it can be easily adapted to other calendar systems.

5.3.7 Other Features

The existence of RSS feeds is a feature that aims on representing the way a web resource publishes information about its updates. It is primarily intended to differentiate between the focused genres and miscellaneous pages.

The CSS rule count is a counter of all CSS rules that are linked to this web resource (cf. [178]). Both internal and external styles are honoured. The reason for this feature is that the focused web genres usually are heavily styled using CSS, whereas some other web genres (e.g. articles) are usually less visually structured and thus contain less rules.

The CSS byte count is a feature that introduces information about the size of CSS data in bytes. It supplements the CSS rule count feature.

Related approaches apply some of the features named above in combination with linguistic features (as presented in section 5.2). However, as LIGD completely relies on the structural properties of HTML documents, linguistic features are ignored altogether and the textual content of a web resource is not analysed.

[3] `http://www.unicode.org/charts/PDF/U0000.pdf`, *retrieved 2011-01-19*

5.4 The Evaluation Corpus

Training a classifier for web genres needs a corpus containing example instances to learn by correctly (manually) classified examples and to validate the selected features, checking whether they are appropriate and distinctive.

In order to evaluate LIGD a corpus is needed that contains samples of all web genres (*blogs*, *forums*, *wikis*) that are taken into account. In order to emulate realistic use conditions, the requirements for this corpus are inclusion of

1. up–to–date, real–world HTML (as opposed to the corpora of other, older approaches that partially only include table–based layout).
2. web resources composed in different languages.
3. content from different authoring applications (e.g. in the case of blogs *Wordpress*[4], *Blogger.com*[5] as well as less–used applications), as they represent and structure content in different ways.

Although some related approaches' corpora (notably Meyer zu Eissen et al. [132] and Santini [168]) are available for research, these do not match the requirements for LIGD, as they are out–dated, feature only English web resources or do not match the selection of web genres. Thus, a novel corpus has been built[6].

Superordinate genre	Sub–genres	Description
Blog	Blog Page (BSP)	The start page of a blog, typically featuring multiple post teasers
	Blog Post (BPP)	The view of a single post with accompanying comments
Forum	Forum Page (FSP)	Typically an overview of topics or threads
	Forum Thread (FTP)	A thread with multiple comments to the lead post
Wiki	Wiki Page (WP)	A wiki page containing one article
Miscellaneous	Miscellaneous Page (MP)	A page that belongs to neither of the web genres above

Table 5.1: Overview of all sampled genres and their respective sub–genres. Note that *Miscellaneous* does not constitute an own genre.

After a preliminary analysis of the targeted web genres, the *blog* and the *forum* genre corpora were split into sub–genres in order to reflect the structural diversity within the different web page types in the web genres themselves. An overview of the resulting web genres can be found in table 5.1. In the following sections the way the web resources have been acquired for each respective web genre is described in detail.

[4] `http://wordpress.org/`, *retrieved 2011-01-10*

[5] `https://www.blogger.com/`, *retrieved 2011-01-10*

[6] However, the Meyer zu Eissen–Corpus is used for an evaluation in section 5.5.5 for showing the limitations of LIGD concerning relatively unstructured web genres.

5.4.1 Blog Pages

There is a considerable structural difference between the start page of a blog and a page containing the specific blog posts, with most of the Blog Start Pages (BSPs) displaying the most recent blog posts (often in an abbreviated or truncated form as a teaser) without comments and additional menus and sidebars, and the typical Blog Post Page (BPP) providing only one single, full blog post with possibly additional comments.

Initially, example BSPs were gathered by extracting the appropriate categories from the ODP[7]. This multilingual web site directory — founded in 1998, then bought by Netscape — follows the Open Content paradigm[8] and is maintained by volunteers that monitor the quality of submitted links, thus avoiding spam sites and broken web links. ODP provides a RDF dump[9] that contains all directory data freely available for download. The Resource Description Framework (RDF) [157] dump was parsed and all web links in categories like `weblog` were selected, resulting in 15,000 instances of BSPs composed in different languages and provided by different authoring applications (e.g. Wordpress, Blogger and others). The respective web resources were downloaded and *exactly one* link was automatically extracted (in order not to skew the representativeness of the BPP corpus) from each page by inspecting the RSS–feed. Finally, these linked HTML documents were downloaded, resulting in 11,800 BPP instances (77% of BSP instances). Some blogs did not have a RSS–feed, thus the BPPs could not be extracted automatically from these pages. Nevertheless, great care was taken to include the major blog engines that were in the BSP corpus.

5.4.2 Forum Pages

Depending on the specific forum application used, there are two distinctively structured page types in a forum. Forum Start Pages (FSPs) are the entry point, giving an overview over all different forums, whereas Forum Thread Pages (FTPs) contain a first post and the following thread of discussion written by different members of this forum.

Scraping the ODP RDF dump for *forum* pages did not prove to be too valuable due to a lack of reliable categorization, therefore FSPs were gathered by scraping the *Big Boards website*[10] (an edited website directory tracking the most active message boards in several languages on the web), ensuring that different forum authoring applications and forums in different languages were contained. Thus, 1,800 FSP instances were obtained. Taking these as starting points with an approach similar to the one taken to extract the BPP corpus, 1,400 FTPs were gathered.

5.4.3 Wiki Pages

Wiki Pages (WPs) most often do not differentiate between a start page and arbitrary wiki pages, thus this genre was not split up. As building a corpus was not possible using ODP data due to lack of an appropriate category system, WPs were obtained by scraping the *Wikiindex* website[11] (3,500 wiki links) and *Wikiservice.at* website[12] (650 wiki links). These two sites provide directories of known wiki com-

[7] `http://www.dmoz.org`, *retrieved 2008-05-23*
[8] `http://www.opencontent.org/opl.shtml`, *retrieved 2011-01-11*
[9] `http://rdf.dmoz.org`, *retrieved 2008-04-12*
[10] `http://rankings.big-boards.com/`, *retrieved 2008-04-20*
[11] `http://www.wikiindex.org/index.php?title=Category:All`, *retrieved 2008-04-23*
[12] `http://www.wikiservice.at/gruender/wiki.cgi?WikiVerzeichnis`, *retrieved 2008-04-28*

munities. After having extracted all linked wiki pages, 3,100 valid WP instances in different languages and provided by different applications were obtained (after having sorted out inaccessible and obviously erroneous pages).

5.4.4 Miscellaneous Pages

In real–world settings, it is not only important to know which of the targeted web genres a web resource follows but also if a web resource belongs to *any* of the genres. Therefore, another class of web resources, Miscellaneous Pages (MPs) that are not contained in any of the three focused web genres, was added to the corpus. Kennedy and Shepherd [100] call this genre *noise*, as it is different to all other genres they take into account. For this collection, 347 web resources were manually sampled belonging to multiple genres based on a list of genres (see appendix section C.2). The obtained web resources are very heterogeneous in length, structure and figure of speech. Additionally, the selected web resources are written in different languages.

These MP instances are not included in the web genre corpus but are added for separate evaluations presented in sections 5.5.4 and 5.5.6.

5.5 Evaluations of Language–Independent Web Genre Detection

Unfortunately, the above–mentioned corpus contained a substantial part of web resources that could not be used in this evaluation. This is due to several reasons:

- Some web resources were still accessible but had moved (e.g. by the domain's owner setting up a new blog engine) or disappeared and the respective "page not found" or "redirect" server headers were not set correctly. Thus, these pages were downloaded and contained in the corpus.
- In some cases, the content of a web resource was not accessible, but instead a page was displayed denoting a server or application error.
- Some web resources contained malicious ECMAScript code.
- Some domains expired after their pages had been added to the ODP and the respective domains were used to host spam. As not all links in the ODP are regularly checked, there was a considerable number of contents of dubious nature.

Thus, all web resources had to be checked manually. As this would have taken a lot of time with the whole corpus, a representative sample of 200 instances per (sub)genre were randomly selected, resulting in a corpus containing 1,000 multilingual instances (for language distributions in sample see table 5.2) in five different genres or sub–genres (cf. table 5.1).

Using this corpus, six different experiments were conducted:

1. Classification using features derived from related work without the pattern features (see section 5.5.1). This section provides an evaluation design and the baseline for the following experiment. Further, this experiment shows that web genre detection on web resources belonging to one of the targeted web genres is feasible and good results can be achieved.
2. Classification with all available features, including the pattern features. This experiment shows the influence of the pattern features on accuracy improvements (see section 5.5.2).
3. Classification only using the pattern features. This experiment serves to show that the novel pattern features contribute to classifying the specific web genres *blog*, *forum* and *wiki* (see section 5.5.3).

Language	Amount	Percentage
English	657	65.70%
German	79	7.90%
French	69	6.90%
Catalan	29	2.90%
Spanish	23	2.30%
Dutch	20	2.00%
Italian	18	1.80%
Portuguese	14	1.40%
Others (incl. Asian languages)	91	9.10%
Total	1,000	100.0%

Table 5.2: Language distribution in corpus sub–sample (without Miscellaneous category). Note that the majority of languages are authored in European languages. This is due to a bias of the underlying data sources.

4. An extended corpus containing not only the focused web genres but also *Miscellaneous Pages (MPs)*, which are arbitrary pages that do not follow one of the respective genres (see section 5.5.4). This is important as it allows drawing conclusions on the applicability of LIGD in real–world settings. In this evaluation, all features are employed.
5. Classification using the web genre corpus from Meyer zu Eissen [132]. This evaluation shows the limitations of this novel approach concerning other, less–structured genres (see section 5.5.5).
6. An extended feature set that contains limited linguistic features for application in real–world scenarios. Practice shows that, even when the language of a web resource is not English, the underlying system that generates the web page contains structural properties like link names and ID/class names that hint towards a certain genre. This experiment shows that taking into account this property boosts the classification significantly (see section 5.5.6).

For the evaluation, the Weka Machine Learning Toolkit [197] was used. Based on the findings of [65], Support Vector Machines with Sequential Minimal Optimization (SMO) [149] were applied for classification.

All classification results were subjected to ten–fold cross validation, meaning that the corpus is partitioned in ten random[13] sub–samples with nine of these being used as training data and the last sub–sample being used as validation data in order to average the classification results, diminish influence of random sample effects and provide a clean train–test–split.

5.5.1 Evaluation without the pattern features

This experiment reflects the non–linguistic features that are used by related work in different combinations and represents the baseline for further improvement.

The primary performance measure is *accuracy*, i.e. the ratio of correctly classified instances for all five genres. As the genres *blog* and *forum* are separated in two structurally different sub–genres, there is additionally the *three–genre accuracy* (in the following abbreviated with *3G accuracy*) which represents the accuracy for the aggregated superordinate genres. Further performance measures are *precision* and *recall*: the *precision* for a class is the number of instances that have been correctly labelled as belonging to the class divided by the total number of elements labelled as belonging to the class (i.e. the sum

[13] Using the default random seed.

of both true positives and false positives, which are instances incorrectly labelled as belonging to the class). *Recall* is defined as the number of the correctly classified instances divided by the total number of instances that actually belong to the class (i.e. the sum of true positives and false negatives that are instances which were not labelled as belonging to that class but should have been). The *F–Measure* is a weighted, harmonic mean of precision and recall.

a	b	c	d	e	← classified as
160	10	15	11	4	a = BSP
17	165	11	0	7	b = BPP
14	6	177	0	3	c = WP
4	0	8	180	8	d = FSP
4	5	6	11	174	e = FTP
0.80	0.89	0.82	0.89	0.88	Precision
0.80	0.82	0.89	0.90	0.87	Recall
0.80	0.86	0.85	0.90	0.88	F–Measure
				85.6%	Accuracy
				90.2%	3G Accuracy

Table 5.3: Confusion Matrix for classification without the pattern features

The confusion matrix — a table showing which instances have been correctly or erroneously classified as which class — in table 5.3 shows that 85.6% accuracy is achieved as the result of SMO classification. Further, one can see that a major source of incorrect classification is the distinction between blog start pages (here labelled as class BSP) and blog post pages (BPP). As these are affiliated with the same superordinate genre (same with FSP and FTP), these results can be merged if detecting only the genre (and not the exact sub–genre) is important, yielding an overall 3G accuracy of 90.2%.

5.5.2 Evaluation using all features

a	b	c	d	e	← classified as
174	14	7	4	1	a = BSP
21	174	2	1	2	b = BPP
15	1	181	0	3	c = WP
10	0	2	180	8	d = FSP
5	4	0	4	187	e = FTP
0.77	0.90	0.94	0.95	0.93	Precision
0.87	0.87	0.91	0.90	0.94	Recall
0.82	0.89	0.92	0.93	0.93	F–Measure
				89.6%	Accuracy
				94.3%	3G Accuracy

Table 5.4: Confusion Matrix for classification using all features with SMO

Table 5.4 shows the confusion matrix of the evaluation result of classification integrating the pattern features in addition. With 89.6%, the accuracy is 4% better than classification results using the features from related work only, if the sub–genres are aggregated, the 3G accuracy gain is 4.1% with 94.3%.

Although the accuracy of the approach using pattern features is performing better, there is a property in the results that is peculiar: When contrasting the classification using only the non–pattern features (cf. section 5.5.1), one can see that especially the precision of detecting the BSP sub–genre is negatively affected by employing the pattern features. Potentially, this is due to the occurrence of large patterns which is one characteristic of BSP. If these large patterns occur in other genres' instances, the classifier may inadvertently identify these instances as BSP.

To determine the statistical significance of the difference of 4% between using or not using the pattern features, the classification was repeated 100 times with ten–fold cross–validation using a different random partition of the dataset (cf. table 5.5). Applying Student's t–test [69] shows that the dataset using the pattern features performs significantly better ($t(0.99; 198) = 69.18$) than the dataset using conventional features only.

	With pattern features	Only non–pattern features
Ten–fold Cross–Validation Tests (n)	100	100
Correctly classified instances (Mean, in %)	89.80	86.05
Standard Deviation	0.393	0.373
Mean Standard Error	0.039	0.037

Table 5.5: Descriptive Statistics of 100 executions of classification with randomized 10–fold cross–validation for datasets using and not using pattern features.

Ranking all features by *Information Gain* [138], measuring how much information a given attribute adds to the distinctive power of the training examples according to their target classification, shows that amongst the 20 most important features are HTML tag frequencies, syntactic URL analysis, link analysis, HTML facets and two of the newly proposed features, `pattern_median` and `pattern_ratio_text` (for a listing of the top 50 features cf. table C.2 in appendix C). This means that the ratio of how much content of a web resource is contained in recurring patterns is meaningful to the web genre of this resource.

5.5.3 Evaluation using only the pattern features

In this evaluation, the quality of classification using only the pattern features is examined. The requirements for the pattern features to perform well is the existence of patterns in the targeted web genre pages. A closer analysis of the 1,000 sample instances shows that in only 23 instances (2.3% of the overall corpus) no patterns can be found, while the average web resource contains about 50 patterns. This means that nearly in all instances patterns can be found and extracted and therefore, the pattern features seem to be valid and applicable to all the focused genres. In figure 5.5, the pattern count distribution is presented. Notably, a quarter of the web resources contain more than 100 patterns, which may seem to be excessive. In many cases, these patterns consist of comments in a forum thread or a blog post. However, as these comments themselves often follow an intricate structure (e.g. having a header containing the author and a footer displaying the timestamp of the comment), also this sub–structure itself can be recognized as a pattern if it is complex enough.

Further, the sample data is classified using only the nine pattern features in order to evaluate the features' expressiveness (see table 5.6). This performs well with 62.0% 3G accuracy for the superordinate genres and 52.1% accuracy for the sub–genres. This shows, that — even if classification based on pattern features *alone* is not performing comparably to classification based on all available features — features representing structural properties of a web resource beyond counting occurrences of HTML elements are

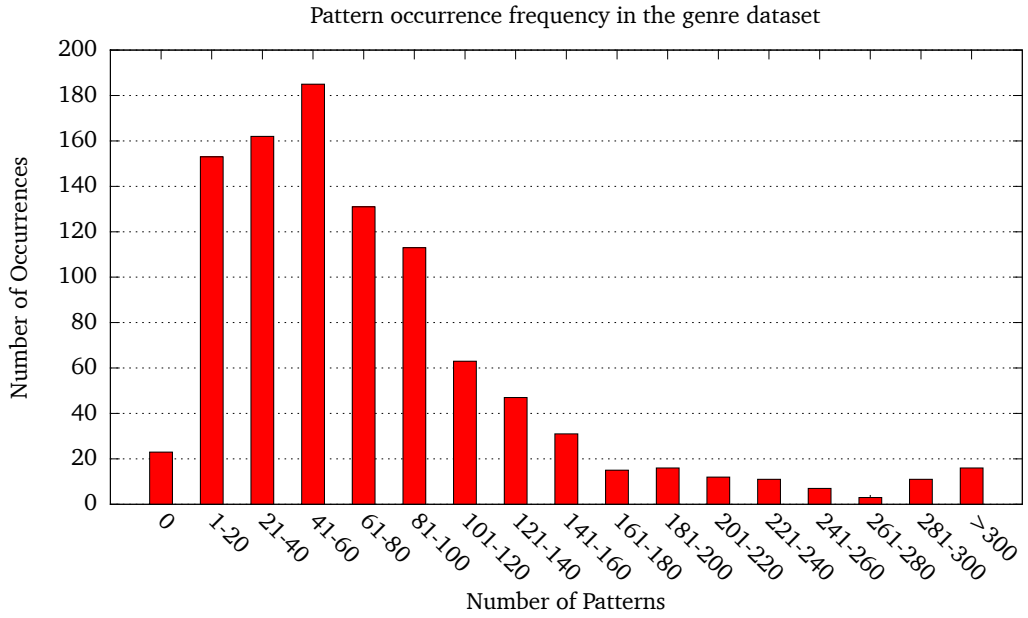


Figure 5.5: Diagram of pattern occurrence frequency in the corpus

a significant addition to other features. Further, it shows that with only nine features, genre detection well beyond the baseline of 20% by guessing or one rule classification is possible.

a	b	c	d	e	← classified as
74	33	36	39	18	a = BSP
15	111	41	16	17	b = BSP
6	64	108	14	8	c = WP
15	9	13	147	16	d = FSP
17	32	35	35	81	e = FTP
0.58	0.47	0.47	0.59	0.58	Precision
0.37	0.56	0.54	0.74	0.41	Recall
0.45	0.49	0.50	0.65	0.48	F–Measure
				52.1%	Accuracy
				62.0%	3G Accuracy

Table 5.6: Confusion Matrix for classification using only pattern features

5.5.4 Evaluation extending the Corpus with arbitrary Web Genres

The confusion matrix of table 5.7 shows the evaluation results of classification using all features. MP, representing arbitrary web genres, is added to the corpus. These MPs are characterized by a strongly heterogeneous appearance and structure, making it hard to detect the differences between MPs and the other genres. The overall accuracy is acceptable with 77.4% correctly classified instances, for only considering the three major genres a 3G accuracy of 79.2% can be achieved. The fact that — despite the challenges that the aforementioned heterogeneity of the web resources introduce — the accuracy has degraded only by about 12% in comparison to the pure–genre evaluation, shows that the proposed features capture the peculiarities of the genres in most cases. However, a notable exception is the WP

genre. The lack of a homogeneous structure in the WPs diminish the discriminative power of structure features and thus the MPs and WPs are easily confused. Besides that, the heterogeneity of the MP class accounts for a higher rate of confusion with other genres, nevertheless the SMO classifier used was able to separate the MP clearly from other genres. This indicates that the presented genre detection is applicable in real–world use cases that involve detection of the targeted web genres.

a	b	c	d	e	f	← classified as
149	4	0	0	3	44	a = BSP
12	160	1	1	0	26	b = BPP
2	5	113	2	1	77	c = WP
2	0	0	171	5	22	d = FSP
3	2	0	3	163	29	e = FTP
15	19	4	13	9	285	f = MP
0.81	0.84	0.96	0.90	0.90	0.59	Precision
0.76	0.80	0.57	0.86	0.82	0.83	Recall
0.78	0.82	0.71	0.88	0.86	0.69	F–Measure
					77.4%	Accuracy
					79.2%	3G Accuracy

Table 5.7: Confusion Matrix for classification with Miscellaneous Page class

5.5.5 Evaluation of Meyer zu Eissen Corpus

a	b	c	d	e	f	g	h	← classified as
84	2	5	12	6	3	10	1	a = *Article*
14	92	1	0	9	5	1	5	b = *Discussion*
1	2	39	3	29	40	13	24	c = *Download*
24	3	14	52	7	14	11	14	d = *Help*
5	1	9	4	115	33	17	20	e = *Linklist*
2	5	19	5	11	83	15	23	f = *non–priv. Portrayal*
14	0	3	6	16	16	71	0	g = *private Portrayal*
2	2	6	1	11	38	3	104	h = *Shop*
0.575	0.86	0.406	0.627	0.564	0.358	0.504	0.545	Precision
0.683	0.724	0.258	0.374	0.564	0.509	0.563	0.623	Recall
0.625	0.786	0.316	0.468	0.564	0.42	0.532	0.581	F–Measure
							53.33%	Accuracy

Table 5.8: Confusion Matrix for classification with Meyer zu Eissen–Corpus (cf. [132])

In order to show the limitations of LIGD, a classification attempt using the Meyer zu Eissen–Corpus and their identified genres [132] is performed using all features. Table 5.8 shows the results. With an overall accuracy of 53.3% (respectively 23.5% using only the pattern features; this is not shown here) the results are sub–standard compared to the results of Meyer zu Eissen et al. (their performance ranges from 60% to 80%).

This is due to several reasons:

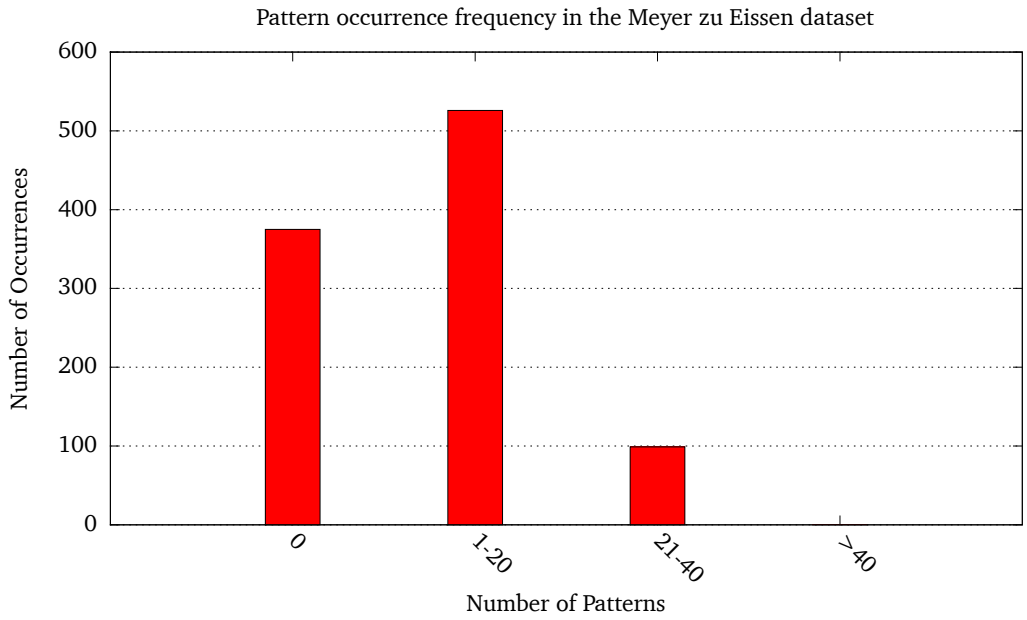


Figure 5.6: Diagram of pattern occurrence frequency (Meyer zu Eissen–Corpus, cf. [132])

1. In contrast to the genres applied in this work, linguistic analysis is more appropriate to these genres and may provide more cues regarding classification. For an example, the genres *non–private* and *private portrayal* differ in style of language used (choice of terms and colloquial language), however, the typical structure of these genres is quite similar.
2. The Meyer zu Eissen–Corpus was built in 2004 and therefore — taking into account how the WWW has evolved in the last few years — is quite out–dated. Only the last few years have seen an increase in use of semantic HTML, where structure and not presentation is focused. Thus, the genres used here are not as structured as the presented pattern detection algorithms require. A closer inspection of the detection results yields that only few patterns can be found at all, more than 30% of all instances do not yield any pattern (see figure 5.6) and 87% of the instances do not have any outer patterns. Therefore, the classification algorithm does not have the necessary data to distinguish the underlying web genres based on these pattern features.

A conclusion from this evaluation is that the presented approach — while being appropriate for the genres *blog*, *wiki* and *forum* — fails on other genres that contain less (hierarchically) structured web resources.

5.5.6 Evaluation using restricted linguistic features

As related work (e.g. [168]) shows, linguistic features boost the performance of web genre detection significantly. However, in a multilingual corpus this is problematic due to the different languages used. Without resorting to an approach that maps genre–typical multilingual feature terms, linguistic features are not applicable to such a corpus. However, if the use of language in a web resource is differentiated between use of a language in its content and use of language in its technical framework, there is a chance to improve the results presented above.

Forum, blog and wiki software often is open source and is therefore developed by a community. Although the developers often are not native English speakers, they collaborate in the lingua franca of

the web, which is English. Consequentially, the source code and the documentation often are written in English. In desktop applications, this is not necessarily visible to the user of a software, but in web applications, some of the internals are exposed through the HTML markup served to the browser. Respectively the classes and IDs of HTML elements (cf. section 4.2) often expose semantic information about the content they markup. For example, a common characteristic that is found in a blog application is that the class names for marking up comments contain the terms `comment`, `post` or `author`. Further, often the URLs of web resources are self–describing towards the genre. This can be utilized to provide *restricted linguistic features* to an approach for detecting the web genre of such a web resource.

The following closed word sets were identified on exemplary web resources that are not in the corpus and subsequently added as feature groups to the feature set described above:

Class and ID names are used to apply style or functionality to HTML tags. Examples for this feature group are *comment*, *header*, *footer*, *navbar*, *blogroll*, *edit* and *archive* that designate certain functionality in a page (e.g. denoting the place for a list of friends' blogs like *blogroll* or marking paragraphs in a wiki as *edit*able).

Domain names often hint to the web genre of a resource. For example, *board* and *forum* are quite distinctive for forums whereas the sub–string *wiki* usually hints towards the web genre being a wiki.

URL sub–strings may indicate the use of a certain web genre. The string *thread*, for example, hints towards forums, whereas the string *post* is usually found in a blog.

Link sub–strings are similar to URL sub–strings, only that they denote the genre a web resource links to instead of designating the web genre of the resource itself. The intention of this feature group is the hypothesis that a blog will more often link to other blog pages than to e.g. forum or wiki pages.

In total, these feature groups add 196 features to the classification, which is approximately 1.2 times the size of the original feature set. As they are not computationally expensive to collect, they represent a light–weight addition to the features in comparison to full–fledged, heavy–weight linguistic features like POS tagging.

a	b	c	d	e	← classified as
190	6	2	2	0	a = BSP
19	177	2	1	1	b = BSP
9	2	188	0	1	c = WP
7	0	2	187	4	d = FSP
6	3	0	4	187	e = FTP
0.82	0.94	0.97	0.96	0.97	Precision
0.95	0.89	0.95	0.94	0.94	Recall
0.88	0.91	0.95	0.95	0.95	F–Measure
				92.9%	Accuracy
				96.2%	3G Accuracy

Table 5.9: Confusion Matrix for Evaluation including selected linguistic features

Table 5.9 presents the confusion matrix that shows that the results are considerable with 92.9% accuracy and 96.2% 3G accuracy. These results are significantly better than the results without the restricted linguistic features ($t(0.99; 198) = 99.13$). On ranking the newly introduced features by Information Gain, especially the link sub–string feature group shows to be promising, supporting the hypothesis that web genres primarily link to other web resources of the same web genres.

If the MP genre is taken into account, the accuracy drops to 87.4% (respectively 89.6% for 3G accuracy). This is still a 10% improvement in comparison with the evaluation presented in section 5.5.4 and shows that LIGD is applicable to real–world scenarios.

5.6 Conclusions

In this chapter, an approach to automatically detect the genre of a web resource has been presented in order to recognize the genres *blog*, *wiki* and *forum*. In ELWMS.KOM this information can be used as metadata, helping learners to create a consistent vocabulary in their resource organization and therefore facilitates both the structuring and the retrieval process. LIGD draws on traditional features from related work, but also introduces novel features that serve to distinguish the web genres by their structure and not the used terminology. The latter base on the hierarchy of the HTML's markup and do not demand knowledge of the language of the HTML's content. Therefore, LIGD is language independent. Further, a corpus has been presented that encompasses 1,000 instances of multilingual resources of the above–mentioned genres, including pages from major providers like Blogspot[14] as well as non–standard and custom blog applications. This shows that LIGD works with different web applications and systems. In the evaluation, accuracies up to 94.3% of correctly classified instances were obtained (89.6% if the exact sub–genres of the superordinate web genres *blog* and *forum* are of interest). Further, these results can be enhanced by introducing linguistic features that depend on the language of the technical scaffold of the respective system, resulting in accuracies up to 96.2%. Additionally, several other evaluations have been performed to show the benefits of LIGD.

LIGD has some advantages over related work, such as the independence of the web resource's language. Further, reasonable results were obtained with only a small set of 144 features. Other approaches — particularly those making use of linguistic analysis — often have several thousand features [118], as they use a possibly large set of closed–class word sets. Thus, the limited number of features in LIGD reduces the computational complexity of the actual classification task significantly.

With the presented accuracies, LIGD is reliable enough to be used in a system like ELWMS.KOM for determining whether a web resource belongs to one of the targeted genres. Though, the use case of LIGD is not restricted to ELWMS.KOM, it has been applied in a Community Mining scenario [61] for identifying a resource's web genre, segmenting the web resource and classifying the content types of the fragments. Thus, in such a setting, it is most useful as a complementary pre–processing step to the segmentation approach presented in chapter 4.

[14] `http://blogspot.com/`, *retrieved 2011-02-17*

6 Supporting Self–Regulated Learning

In self–directed RBL settings using web resources, learners usually are not guided by a teacher or a tutor. Further, as web resources usually are not intended to be used for learning (e.g. weblog posts, wiki articles or community pages), they are not didactically structured and thus rarely provide the guidance that learners need. Additionally, the availability of a dedicated LO that covers the learner's specific information need cannot be guaranteed. Hence, an application like ELWMS.KOM that aims at supporting RBL needs to substitute this lack of direction by enabling the learners themselves to assume the role of the organizer of their learning processes. This involves supporting setting goals, planning the learning process, self–monitoring and reflection, and eventually modification of a sub–optimal process step. Such a support needs to affect all processes of RBL that are executed in the personal context of a learner (cf. figure 6.1).

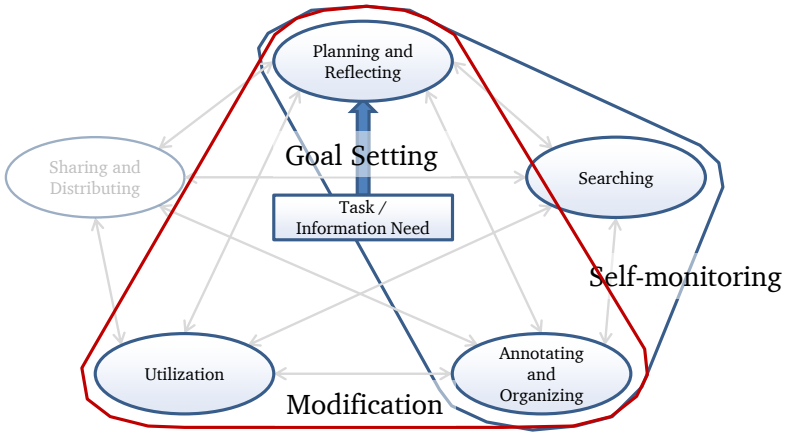


Figure 6.1: Supporting principles of Self–Regulated Learning benefits all processes of Resource–Based Learning in the learner's personal context.

The theory of Self–Regulated Learning provides a framework for giving exactly this support, postulating that learners have to execute the metacognitive processes *setting learning goals*, *planning and monitoring their learning process* and finally *reflecting* on it in order to readjust their procedure for the next learning episodes. This chapter describes an extension to ELWMS.KOM that accommodates principles of Self–Regulated Learning (SRL) by supporting the above–mentioned learner processes.

6.1 Introduction

Major challenges for self–directed learners consist of stating their information needs, formulating search queries, estimating relevance of found resources, filtering irrelevant resources and keeping track of the state of the search process, i.e. monitoring their progress [7, 13]. These processes require high learner's competencies of self–organization and self–motivation, as a deep information search in the context of learning is not trivial. These processes are covered by the theory of Self–Regulated Learning (SRL). Central to this theory is the notion that learning is a process that is self–directed and needs regulation on the learner's side [6].

In the context of this thesis, RBL encompasses this style of learning. As shown in chapter 2, self–directed learners usually identify their information need autonomously and proceed to cover relevant information by searching on the web or dedicated digital libraries. Thus, SRL is applicable on learning settings like the presented one and should be supported in such a self–directed learning process.

6.1.1 Structure of this Chapter

In this chapter, additions to ELWMS.KOM that address the above–mentioned challenges are presented. Section 6.2 presents a basic overview of the theory of SRL that adequately reflects this self–directed process of learning with web resources. Further, the term *scaffolds* that denotes support of this process is explicated. The design and implementation of additions to ELWMS.KOM that enable learners to set learning goals prior to internet search and assign relevant web resources to these goals is given in section 6.3. The goal–setting component has been implemented for ELWMS.KOM that is an add–on for the web browser Firefox, as web browsers are the gateway to most information on the web. Section 6.4 presents two studies and evaluations of ELWMS.KOM showing the benefits of supporting the process phases of SRL and section 6.5 concludes with a short summary and an outlook.

6.2 Self–Regulated Learning and Scaffolds

Self–directed Resource–Based Learning with web resources is a process that involves Self–Regulated Learning. As such it is quite demanding for learners: they have to *plan*, *monitor and regulate* and *reflect and modify* their learning process in order to reduce disorientation and enhance quality of their learning achievements [6, 173, 37, 155]. In the following, particularities of this self–regulated way of learning and possibilities to support it using so–called *scaffolds* are presented.

6.2.1 Self–Regulated Learning

It has been shown that supporting learners to conduct the processes mentioned above can in general improve the learning experience as well as the outcome [173] (e.g. by providing training or instructing learners to write a learning diary). Specifically, for learning scenarios using web resources, supporting SRL has shown to improve learners' understanding and conceptual knowledge of a topic [6, 5].

Central to the theory of SRL is the notion that learning is a process that is self–directed and needs regulation on the learner's side. Therefore, it represents a specialization of SDL that focuses on the psychological processes that are executed by a self–directed learner. Theories of SRL focus on the *personal* learning process, dealing with the goal–setting process and regulation of goal–directed acting, emotions and cognition. The term *regulation* is borrowed from cybernetics, where it denotes control that is geared to frequent measurements of an as–is state and adapts itself accordingly. Thus, *self–regulation* is the process of an individual monitoring itself and thus adapting its acting towards a previously set goal. Bandura [10] defines self–regulation as the ability to actively influence internal sources of acting and experiencing like thoughts, motivation, volition and feelings. Further, he states that self–regulation "operates through a set of sub–functions that must be developed and mobilized for self–directed change", where these sub–functions are self–monitoring, judgements of one's behaviour in relation to personal goals and environmental circumstances and affective self–reaction. SRL transfers this notion of self–regulation into the domain of learning.

According to Boekarts [25], three different systems have to be regulated in order to learn in a self–directed manner (see figure 6.2). The *cognitive system* is performing task editing strategies, so the learner

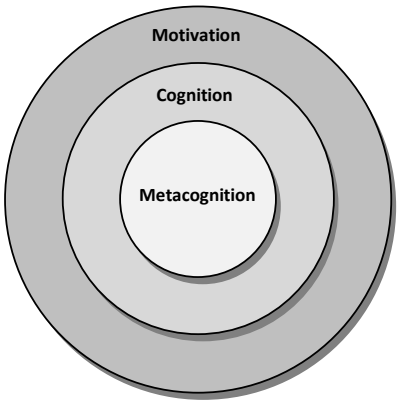


Figure 6.2: The three different systems that have to be regulated for self–directed learning (cf. [25]).

will choose a strategy that she deems to be effective and efficient for a certain task. For example, a learner who is searching for information on the Internet has to think about search query terms that are likely to lead to success, i.e. relevant resources. In the *motivational system*, the learner regulates her volitional and motivational state, so that she will for example overcome procrastination and start a learning episode or cope with obstacles appropriately. Finally, in the *metacognitive system*, the learner sets learning goals, devises plans and strategies for executing the actual learning process, monitors her progress on her actions, re–adjusts them if necessary and reflects on her achievements, eventually leading to forming of strategies to enhance her learning. Boekarts states that for attaining a successful learning process, all of those three systems have to be regulated. A failure in one system cannot be compensated for by the others.

The theory SRL focuses on the cognitive, motivational and metacognitive processes of learning, thus it intends to describe only a learner's personal scope of learning. Cooperative and collaborative learning settings are not targeted by this theory.

As the focus of this chapter is on metacognitive processes, subsequently only processes are considered that are executed in the metacognitive system.

Schmitz and Wiese [173] and Bannert et al. [13] partition the learning process in three phases: before learning (*pre–action phase*), during learning (*action phase*) and after learning (*post–action phase*). In each of these phases, there are processes that are executed in each of the three systems in order to regulate the learning process.

Respectively, the different metacognitive processes performed in each respective phase (see figure 6.3) are beneficial for a successful learning process: Before learning, the learner performs goal–setting and planning, whereas while learning, the progress and course of actions are monitored and —- if necessary — adapted to the current state of the learning process. Finally, after having learned, reflection processes are executed in order to modify and optimize future learning processes.

Benz et al. [20] map the processes described above to learning episodes of different temporal granularity. For example, a learner who searches for a relevant fact on the Web performs a rather fine–granular learning episode. The learner sets her desired search goals, plans and monitors her search process and finally evaluates, whether her learning goals have been met within seconds or minutes. However, a learner working on a bigger project (e.g. a homework, scientific paper or thesis) usually plans her approach, monitors and evaluates her process over several weeks. Still, a project will consist of several smaller (possibly related) learning episodes that are executed in the context of the project.

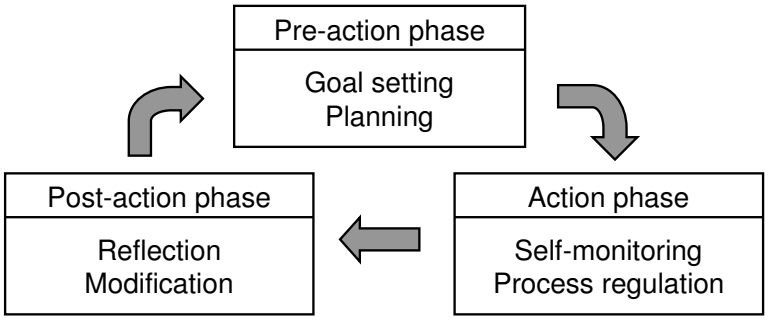


Figure 6.3: The three phases and accompanying metacognitive processes in SRL (cf. [173, 13, 19]). After the post–action phase, the learning process is reflected on and the next learning processes are modified in order to improve their execution.

6.2.2 Goal–Setting and –Orientation

Goal–setting is an essential part of the theory of SRL. Latham and Locke [114] state that learners benefit from setting goals motivationally and improve their learning performance. They define properties, moderators and mechanisms of goal-setting strategies, stating the importance of setting specific and challenging (but not impossible to attain) goals. Further, they state that an important factor of a successful learning episode is the existence of personal attachment to the set goals. According to Latham and Locke, the mechanisms that lead to performance increase are the orientation of acting towards the goal, intensity of those acts, perseverance while attaining a goal and applying goal–directed strategies.

Further, Latham and Locke state that appropriate feedback is essential for the benefits of goals, as it allows a learner to adapt their attitude and attachment towards the goal. Latham and Locke show that a specific goal simplifies the learner's identification and monitoring of the behaviour necessary for goal attainment.

Bandura and Schunk [11] differentiate between proximal and distal goals. Proximal goals are attainable in an immediate or near–term time frame and tend to be very specific. Distal goals, however, are set to be achieved in the distant future and tend to be less specific. Bandura and Schunk state that distal goals are ideally complemented by proximal goals, as these provide a roadmap to the distal goals and thus the learning processes result in improved performance. Therefore, it is important to support learners in the process of attaining the set proximal goals [175].

6.2.3 Scaffolds

Vygotsky [194] introduces the term *scaffolding* as a "guidance provided in a learning setting to assist students with attaining levels of understanding impossible for them to achieve without external support". Thus, scaffolds can be seen as learning aids that help learners to execute qualitative learning processes in order to achieve better learning results. In the long term, scaffolds should be designed to advance competencies, thus learners will not be dependent on the scaffolds anymore. According to Azevedo and Hadwyn [7], scaffolding in computer–based learning environments may support a range of instructional targets:

- Learning domain knowledge. This corresponds to the cognitive system (e.g. by learning concepts and procedures).

- Learning about one's own learning. This is often called *meta–learning* and is clearly regulated by the metacognitive system. Further, metacognitive self–regulated processes are often allocated here.
- Learning about using the computer–based learning environment. This means that a learner has to attain a certain level of expertise in order to use the tools at hand efficiently and effectively.
- Learning how to adapt to a particular instructional context. For example, a learner can be supported by encouraging her to engage in adaptive help–seeking.

According to Friedrich and Mandl [76], scaffolds can be implemented both *directly* and *indirectly*. Direct scaffolds communicate instructions (so–called *prompts*) that ask the learner to carry out a certain learning action. For example, instructing learners to set learning goals before starting to learn is a *direct scaffold*. *Indirect scaffolds* can be implemented by design of a learning environment, so that the learner has the possibility to use certain supporting functionalities if required. For example, providing goal–setting functionality without a dedicated prompt to use it can be seen as an indirect scaffold.

The theory of SRL postulates specific processes that contribute towards a high–quality learning process. The concept of scaffolding defines and describes different possibilities to realize learner supports. Combining both approaches, learning processes can be assisted and supported according to the presented theoretical principles [175].

6.2.4 Supporting Self–Regulated Learning in Resource–Based Learning Scenarios

Multiple researchers have evaluated the effects of RBL using web resources under the pretext of the theory of SRL. In this section, a few selected works that support metacognitive activities and goal setting are presented and discussed.

Bannert et al. [13] present a study on support of metacognitive processes in SRL. Before learning, they provide students with a "metacognitive support device" (basically a training on how to employ metacognitive skills during learning that was presented as hypertext). This training presents specific questions for each metacognitive activity, e.g. "What do I actually want to learn?", "Do I remember and understand the topics I have learned?" or "Did I reach my learning goals?". Additionally, a comprehensive exercise is included to give students the possibility to immediately apply these activities. During learning, the students in the experimental group were instructed to apply these metacognitive activities by being given a paper–based prompt in form of a diagram visualizing the metacognitive activities taught before. Bannert et al. show that the experimental group that received the training and the prompt achieved significantly better results in transfer performance, although a performance increase could not be shown for knowledge and recall tests. This result is in accordance with the findings of a prior study [12], where the participants learned using materials in a closed hypertext environment. Thus, the results of this study shows that supporting metacognitive prompts turns out to be promising. The learning material in [13], however, consists of a linear text that is paper–based, only the training was presented as hypertext. Bannert et al. state that they expect a performance increase if metacognitive processes are supported in more complex learning scenarios, e.g. when learning with open–ended, non–linear hypermedia environments.

Biswas et al. [24, 23] present a teachable agent system called "Betty's Brain" that combines learning by teaching and self–regulation strategies to promote deep learning and understanding. By teaching the virtual agent Betty important concepts and their relations about a knowledge domain, learners have to explicate their knowledge in a concept map so that the agent may infer correctly over this knowledge. Biswas et al. claim that this way of learning provides engagement and motivation to learners by increasing social interactions with the system. In order to support novice learners, they provide cognitive and metacognitive scaffolds that prompt the learners to set goals, reflect on their current knowledge and

performance, and allow testing the current progress of Betty. Biswas et al. [24] evaluated the system by comparing the experimental group being supported with SRL scaffolds to two control groups using only the agent without SRL support varying on the possibility of interacting with the teachable agent. They demonstrate the positive effects of SRL strategies in terms of understanding and transfer performance.

Salmerón et al. [166] state that the need to select appropriate hyperlinks during a search process requires students to self–regulate their search processes to a degree far greater than in regular linear text. They focus on predicting the reading order of interlinked hypertext documents based on different individual factors like the difficulty of the learning goal, prior knowledge, use of learning strategies and comprehension calibration. In two studies, they evaluate the hypothesis that there is a strong relation between SRL and link selection strategies, two factors that they assume to have a robust impact on comprehension. Salmerón et al. show that *efficient* learners regulate their hyperlink selection strategies in order to optimize their learning processes. Further, they conclude that hypertext comprehension is constructed from a combination of self–regulation and different link selection strategies, with the difficulty of the set goal being an important indicator of comprehension. However, the evaluation setting presented in [166] is a simplified hypertext system which consists of a multi–section text in which readers are only able to choose between two further hyperlinks and so restrict the user's link selection behaviour. Thus, it is not a realistic hypertext setting, failing to represent the complexity of typical web pages that often encompass multiple contexts, links or media.

6.3 ELWMS.KOM additions supporting Self–Regulated Learning using Web Resources

In order to successfully execute SRL processes, a certain routine on the side of the learner has to be promoted. Thus, the competences and processes that are needed for an effective self–regulation are usually conveyed in specific trainings. However, trainings (distributed as a WBT or given by a personal coach) are mostly designed to help a learner to getting started and rarely accompany the whole learning process. Therefore, a systemic approach has the advantage that it is available over a longer period and during the actual learning process.

Thus, in this section, a concept of additions to ELWMS.KOM (cf. chapter 2.5) for internet search is derived from the presented theoretical principles of SRL and the implementation is presented. The central additional component of ELWMS.KOM is a goal–management component. Learners can enter goals, organize them into goal hierarchies (setting super– and sub–goals), move them via *drag and drop* and attach found resources relevant to the respective goals. Each goal can have an arbitrary number of sub–goals and resources, organizing everything in a tree structure with exactly one super–goal —- analogue to the directory structure of a common file system.

For the two studies presented in section 6.4, two different versions of ELWMS.KOM were implemented based on the requirements of the respective study.

6.3.1 Conceptualization

The goal–management component is an addition to ELWMS.KOM that partitions the learning process into the three phases *before learning*, *while learning* and *after learning* (cf. [173]). A focus is set on the metacognitive processes of goal–setting, planning, monitoring, regulating and finally reflecting and modification of the learning process. In ELWMS.KOM's base form, the scaffolds that support those processes are implemented *indirectly*, which means that the learner is not instructed to take direct action, but she may choose to use the functionality if she sees the need to (cf. [76]). Before beginning with the internet

search, the learner chooses a goal–directed approach and plans her course of actions in the learning process. For example, if a learner has the task to search for information about the topic *Classical antiquity*, she may begin to structure her approach with the goals "I need to get a general idea about the ancient Rome" and "I need an overview of the ancient Greece". Each goal can be further subdivided into specific sub–goals, e.g. the ancient Rome may contain the sub–goals *Roman Republic* and *First Triumvirate*. This way, the learner organizes her search goals into a goal hierarchy (cf. figure 6.4). Hence, ELWMS.KOM supports processes of goal–setting and planning.

During the learning process the learner may attach found information in web resources to the set goals and rate their relevance for the respective goal. As the learner's information need often is quite specific, just storing a whole web resource is usually not enough. Instead, the possibility to extract only the relevant part of the information is more target–oriented towards the real learning goal. Thus, the selected fragment (called *snippet*) of an imported web resource is stored in the goal's metadata; learners can access that relevant information later without having to access the original web page. Monitoring the learning process is supported by multiple scaffolds, e.g. setting the progress of attainment of a certain goal and displaying the goal hierarchy in combination with the already found web resources. Both stimulate the learner to contemplate where in the learning process she is right now, which goals she has already achieved and what goals are still open. In order not to lose focus on the goal the learner is following right now, in the 2^{nd} version of ELWMS.KOM it is possible for her to activate one goal at a time. This goal is displayed prominently, giving a reminder not to go astray and antagonizing the well–known *lost–in–hyperspace* phenomenon (experiencing disorientation due to information overload and aimlessly following hyperlinks [55]). Further, all goals and found resources can be displayed as a knowledge network (see figure 6.5) and an outline displaying all goals and resources. This enables the learner to reflect on already found information and the current course of action. Is the learner aware of her inefficient advance, she may alter her search behaviour according to her current situation — for example by defining new goals, re–structuring her goal hierarchy or focusing on other goals that are more promising at the moment. Hence, during the search the processes of monitoring and regulation are supported.

After learning, the learner has the choice between different alternatives of visualizing all goals and resources: the goal hierarchy, the knowledge network and the outline. However, the theory of SRL differentiates between the monitoring and regulation processes mentioned above and the processes of reflection and modification, as these occur after having finished the search process. Here, the visualizations enable learners to reflect on the finished learning episode, both from the view of the results and the taken approach. Thus, if the learner decides to optimize her approach based on her reflection, processes for modifying the approach are executed.

Metacognitive Processes	Supporting function in ELWMS.KOM
Goal–setting & Planning	Creating goals, structuring the goal hierarchy
Monitoring & Regulation	Setting the progress of a goal, displaying the goal hierarchy in combination with already found goals as a knowledge network or a list, activate a currently followed goal, scaffolds prompting learners to plan, monitor and reflect
Reflection & Modification	Different visualizations of content, prompt to reflect on found resources

Table 6.1: An overview of metacognitive processes and the supporting functions in ELWMS.KOM

Table 6.1 briefly summarizes the supported metacognitive processes and the associated supporting functionality in ELWMS.KOM.

6.3.2 Technical Foundations and Implementation

Searching and learning using web resources mostly takes place in the web browser, as most web resources are represented as HTML markup. The browser is a virtual window to the Internet, downloading and rendering web resources and displaying them to the learner. Therefore, ELWMS.KOM has been implemented as an add–on to the popular open source web browser Firefox (cf. chapter 2.5).

Due to portability and extensibility reasons the core functionality has been realized in a Java[1] applet. Data transmission between Firefox and the applet is performed via an ECMAScript[2] interface that both orchestrates the data flow and forwards user interaction within Firefox or the web resource to the applet. The graphical user interface and data storage has been implemented in Java. Applets as a technology were chosen as they allow integration in HTML as well as in XUL[3], the Firefox–specific XML dialect for creating graphical user interfaces.

The metadata of goals consist of a title, a description (which may serve to outline a course of actions or additional information) and the level of progress (with the stages *not started*, *25%*, *50%*, *75%* and *finished*). This level of progress can be set by the learner to keep an overview of her open and finished goals. Further, goals can be tagged (i.e. attaching freely chosen key words) for organization and display in the knowledge network. For long–term learning episodes, further functionalities that provide adequate archiving and retrieval would be necessary, but in this work this has been neglected because only short–term learning episodes are focused.

The web resources are inserted into goals by use of the *import* functionality, similar to the process of bookmarking in a web browser. Similar to goals, resources have a title, a description, a relevance rating and tags. The description encompasses the snippet of a web resource the learner has selected and / or arbitrary text. Rating the relevance of a resource or the snippet with the stages *not rated*, *not relevant*, *a little relevant* and *relevant* is possible as well.

The goal–management component for ELWMS.KOM is displayed in the browser's sidebar. Its user interface shows an overview of the current goal hierarchy and resources (see figure 6.4). Alternative representations of goals and resources may be used, e.g. a display of the goal hierarchy as a knowledge network (figure 6.5). While browsing, web resources can be imported into the goal tree at the current selection. Both goals and resources may be edited and reorganized later–on.

[1] `http://www.oracle.com/technetwork/java/index.html`, *retrieved 2011-01-12*
[2] `http://www.ecmascript.org/`, *retrieved 2010-11-30*
[3] `https://developer.mozilla.org/En/XUL`, *retrieved 2010-11-30*

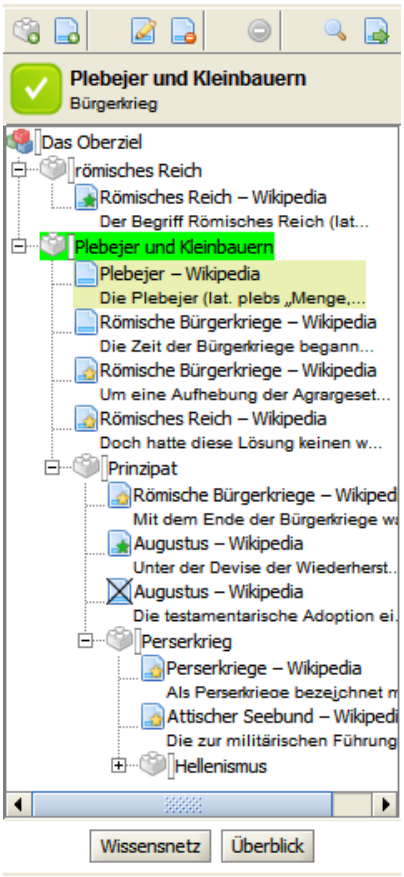


Figure 6.4: Screenshot of the 2^{nd} version of ELWMS.KOM in the sidebar of Firefox. The goal hierarchy is shown with the currently attained goal prominently presented at the top. The brick icons denote goals, the page icons denote web resources. In this example, the learner has adopted the titles of the web resources as descriptors of the web resources, thus the descriptors reflect their source with the Wikipedia page title.

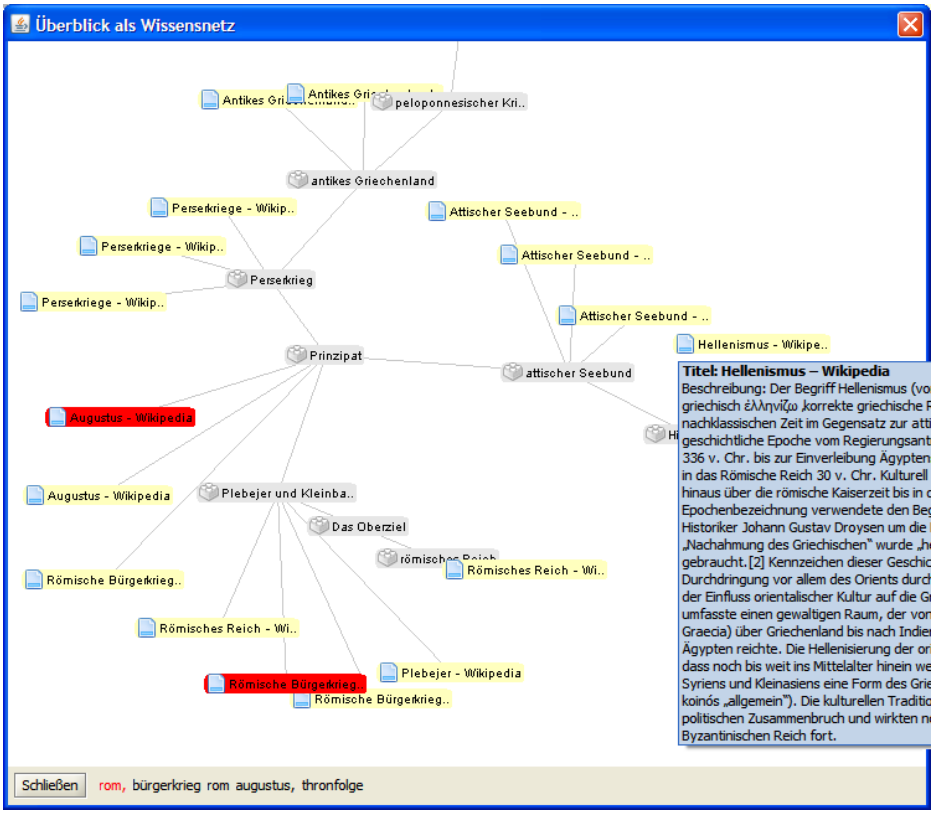


Figure 6.5: Screenshot of a Knowledge Network built by connecting Resources and Goals via Tags.

6.4 Two User Studies

For evaluating the goal–setting component of ELWMS.KOM, two user studies were performed[4]. The first study served as an exploratory setting for determining adequate support possibilities for SRL processes, whereas the second study built on the findings of the first study, focusing on specific aspects of interest and on improving ELWMS.KOM. In the following, the commonalities of both studies will be briefly described, moving on to the peculiarities and conclusions of the first study and eventually focusing on the second study.

6.4.1 Commonalities of Both Studies

Both studies presented in the following sections focus on evaluating the effects of supporting metacognitive SRL processes while learning using web resources. As opposed to related approaches, a realistic, open environment comparable to the web is targeted. However, a study design that allows browsing the whole Web for information makes the estimation of effort on the side of the learner difficult. Thus, it was decided to limit the scope of the hypertext system used to browse for learning materials to the German Wikipedia[5]. Wikipedia is a collaboratively created online encyclopedia that features articles about a wide range of topics. The articles are strongly interconnected by wikilinks which consist of hyperlinks to other

[4] These studies were performed in a joint project with the department of Psychology of the Technische Universität Darmstadt as part of the Research Training Group on *Feedback Based Quality Management in eLearning*. For further information, see the publications [175, 174, 19].

[5] `http://de.wikipedia.org/`, *retrieved 2010-09-12*

articles concerning a related topic. As the article link structure of Wikipedia has similar properties as the overall web [207], it is well suited for such a task.

For both studies, the topic *Classical Antiquity* has been chosen as the subject of the learning task, as the participants were expected to have little prior knowledge about it. This topic is well–covered in the German Wikipedia, there is a wide range of articles about *Ancient Rome* and *Ancient Greece* that are aggregated in a portal article[6]. From there, the majority of all needed information is directly linked. Examples for test questions are "Which event led to the end of the Roman Kingdom?" and "Arrange the historic periods of the Ancient Rome in the order of their temporal occurrence". The questions were given as a pre–test before learning so that the participants knew what was expected of them.

During the learning phase, several data about the learning process were logged. First, the screen content of the participants was recorded using a screen–capture program[7] and, additionally, the click path of all opened learning resources was logged on the client side (i.e. the respective URLs, a time stamp and the id of the browser tab where the resource was opened). Further, in the evaluation groups using ELWMS.KOM, the creation and changes of the set goals and used resources were tracked and stored in an XML–based log file. In both studies, the actual learning phase was 45 minutes, as this research focuses on short–term learning episodes.

6.4.2 Exploratory Study

The first study [175] focused on the research question how different tools used in learning processes affect the way learners execute their learning processes using web resources. Further, the question how explicit prompts can be given in order to initiate goal–setting, planning and reflection processes [21] was targeted. 64 participants (all psychology bachelor students in their first two semesters between the age of 19 and 28 years, 76.6% being female due to the field of study) were asked to take part in a study that targeted learning using the German Wikipedia for 45 minutes. Four different treatment groups were formed by random assignment:

- *Control Group 1* ($CG_1 1$, $n = 16$) used only the tools the web browser Firefox provides. Participants were allowed to bookmark found resources.
- *Control Group 2* ($CG_1 2$, $n = 18$) used pen and paper as the means to persist their findings. The participants of this group were allowed to take notes during the search.
- *Treatment Group 1* ($TG_1 1$, $n = 15$) used ELWMS.KOM without specific instructions how to proceed. Thus, this group represents a realization of indirect scaffolds as defined in section 6.2.
- *Treatment Group 2* ($TG_1 2$, $n = 15$) used ELWMS.KOM and were asked to execute metacognitive processes by metacognitive prompts in the learning phase. Before learning, participants of this group were prompted to set their search goals. After the learning phase, the participants were instructed to reflect over the found resources and their created goals for five minutes.

Study Design

The study followed a design that is briefly outlined in the following: Before the participants started with the learning episode, they were given a pre–test questionnaire that collected basic demographic data, computer literacy (based on their self–conceptualization), learning competencies (i.e. the competencies

[6] `http://de.wikipedia.org/Portal:Antike`, *retrieved 2010-12-01*

[7] Free Version of Camtasia Studio 3.1.2, `http://www.techsmith.com/camtasia/`, *retrieved 2006-06-15*

to plan and structure their learning processes based on items from [195]) and their current motivation and confidence in their learning competencies. Further, their emotional traits were measured according to PANAS [111], a standardized questionnaire aiming at measuring positive and negative emotions. Then, depending on their experimental conditions, they were given a five–minute introduction in either ELWMS.KOM or Firefox. Before learning, all participants processed a performance test containing 30 multiple–choice questions about the topic *Classical Antiquity*. They were suggested that the same test would be given again in a post–test, enabling competent learners to identify knowledge gaps in the pre–test and reformulate these into learning goals. In the learning phase, all groups were given hints about the remaining learning time after 25 and 40 minutes. After learning, the performance test was handed out again, followed by a questionnaire asking about the participants' approach during learning, the current motivation and emotions again.

Selected Results

In total, during the learning phase all 64 participants viewed 242 unique web resources, that is on average 17.23 resources per participant (when counting repeated viewing, 22.38 web resources were opened). An important observation was that a majority of goals was rather formulated as a *topic* than as real goals (e.g. "I want to get an overview of the topic *Antiquity*"). On inquiry, participants stated that this was due to the given task and time constraints and that the user interface of ELWMS.KOM is only displaying the first words of long goal names.

Selected Group differences

The groups were contrasted over the whole learning process based on different variables gained by analysing the log files and questionnaires. Two group comparisons were executed by applying Student's t–test [69], for the comparison of more than two groups, one way ANOVAs [69] (Analysis of Variance between groups, comparing group means with each other) were used. Table 6.2 shows selected significant differences between the groups. All significance values in the following have been computed one–tailed.

Variable	Contrasted Groups	Student's t–test
Restructure Goal$^{\alpha}$	$TG_1 2 < TG_1 1$	$M_{TG_1 1} = 7.2; M_{TG_1 2} = 3.87; p < .05^{*}$
Revisited Resources$^{\alpha}$	$TG_1 2 > TG_1 1$	$M_{TG_1 1} = 0.2; M_{TG_1 2} = 1.0; p < .05^{*}$
Variable	**Contrasted Groups**	**ANOVA**
Opened resources$^{\alpha}$	$TG_1 1 \cap TG_1 2 > CG_1 1 \cap CG_1 2$	$F(3/60) = 3.65, p < .05^{*}$
Opened images$^{\alpha}$	$TG_1 1 \cap TG_1 2 > CG_1 1 \cap CG_1 2$	$F(3/60) = 1.71, p < .05^{*}$
PANAS "active"$^{\beta}$	$TG_1 1 \cap TG_1 2 > CG_1 1 \cap CG_1 2$	$F(3/60) = 3.19, p < .05^{*}$
PANAS "determined"$^{\beta}$	$TG_1 1 \cap TG_1 2 > CG_1 1 \cap CG_1 2$	$F(3/60) = 4.60, p = .01^{**}$

Table 6.2: 1^{st} Study: Selected Group differences based on log files and questionnaires. F =F–value, p =niveau of significance, M =mean value. * denotes significance ($p < .05$), ** denotes strong significance ($p < .01$). Variables marked with $^{\alpha}$ designate data obtained from logfiles, variables marked with $^{\beta}$ were collected in questionnaires.

Performance was, as expected due to the short learning episode, not significantly different between the groups. However, as $TG_1 2$ "lost" five minutes due to the metacognitive prompts, they had effectively less time for learning. Participants of groups using ELWMS.KOM ($TG_1 1$ and $TG_1 2$) browsed significantly more web resources and opened more images that were partly relevant for understanding the continuity of the learning materials. Further, $TG_1 1$ and $TG_1 2$ benefited from using the goal–setting component

emotionally and motivationally: in comparison to the control groups, they felt they executed their search process in a more active and determined way.

The treatment groups differed significantly on how web resources that were already assigned to goals were handled. $TG_1 2$ changed the already persisted web resources less often, as they already had adequately planned their approach before learning. Thus, they already had a goal to attain and progressed more target–oriented. Further, they revisited their already persisted web resources more often and thus showed to execute reflective processes at the end of the learning phase.

Selected Correlations

Participants that were initially motivated showed a higher willingness to use the goal–setting component intensively and to delve into the learning process. This can be seen in the correlation of the self–declared motivation in the pre–test and the number of followed image links, the number of set goals and their editing and usage of the offered representation of the goal structure as a knowledge network (for correlation values, see table 6.3).

Variable 1	Variable 2	Correlation r
Opened images$^{\alpha}$	Motivation pre–test$^{\beta}$	.252*
Number of set goals$^{\alpha}$		.221*
Edit actions on goals$^{\alpha}$		.226*
Browsing knowledge net$^{\alpha}$		.259*
Search literacy$^{\beta}$	Opened resources$^{\alpha}$	.416*
	Deleted resources$^{\alpha}$	.385*
Computer literacy$^{\beta}$	Opened resources$^{\alpha}$	.525**
	Edited resources$^{\alpha}$	.346**
Restructure goal$^{\alpha}$	PANAS mean negative$^{\beta}$	.353*
	PANAS “confused”$^{\beta}$	.590**
Edited resources$^{\alpha}$	PANAS mean positive$^{\beta}$	.326*
Further usage$^{\beta}$	Motivation pre–test / post–test$^{\beta}$	.343*, .401*
	Revisited resources$^{\alpha}$	−.375*
	Restructure goal$^{\alpha}$	.343*

Table 6.3: 1^{st} Study: Selected significant correlations between variables, * denotes significance, ** denotes strong significance. Variables marked with $^{\alpha}$ designate data gained from logfiles, variables marked with $^{\beta}$ were collected in questionnaires.

The higher the self–perceived search and computer literacy, the more learners were able to filter irrelevant resources and reflect on their already found resources. Additionally, the already saved resources were re–opened and used more often.

If the participants had created a realistic goal structure before having started to learn, they did not have to restructure their goals in the following. At the same time, less negative emotions occurred compared to participants having restructured their goals more often. Especially the feeling of disorientation occurred less often with participants having adequately planned their learning process before–hand. Thus, the planning and pre–arrangement of a realistic goal structure before executing the search using ELWMS.KOM led to a more positive experience of the learning process. Further, the functionality of adapting the resources to the current state of the learning process by stating the relevance of a re-

source and editing the resource's snippet was perceived as useful. Thus, positive emotions accompany the number of resources that were edited, commented and tagged[8].

The less resources the participants collected and the more need they saw to restructure their goal hierarchy, the more the participants stated in the questionnaire that they would like to use ELWMS.KOM or a comparable goal–setting tool in further learning episodes. This may hint that especially learners that do not automatically execute self–regulated processes perceive the need of getting support for self–directed learning using web resources. Furthermore, motivation before and after the learning phase correlates with the desire for using a comparable goal–setting tool in future learning episodes.

Conclusions from Exploratory Study and Implications for Second Study

In conclusion, the presented scaffolds did influence the learning processes, although an impact on the learning performance could not be shown. This was expected due to the short learning episodes, because planning, monitoring and reflecting continuously are expected to have an impact in longer learning settings. Further, in the short time span that the participants learned, the participants could contain all found facts in short–term memory, and therefore the advantage of an elaborate organization of goals and resources was not of primary importance for this learning setting.

Still, other several issues with the study design were encountered. First, the goal was to emulate "realistic" environments for the learners, i.e. forming a control group learning using bookmark functionality and a pen and paper group. Therefore, the groups were not comparable in all accounts and that might have influenced the learning outcomes. For example, the pen and paper group $CG_1 2$ did not have to learn using a new tool and could quickly outline information and set relations between content that was not possible for the other groups. Additionally, the bookmark group $CG_1 1$ was lacking the possibility to save web resource snippets, thus participants had to bookmark the whole page — which many participants thought to be futile, thus not using this functionality at all. Another issue was that the assumption was made that the students had no prior knowledge about the topic *Classical Antiquity* without proving it. Eventually, the groups using the goal–management component were only briefly trained to using it before learning. This means that computer literacy and experience in using comparable tools had a strong influence on the way students were able to handle ELWMS.KOM.

Thus, the study design and some aspects of the goal–management component were redesigned in order to further improve the evaluative quality of the study.

6.4.3 The Second Study — Application of Metacognitive Scaffolds

In the second study, the study design was optimized and a somewhat different scope was chosen. First, sufficient training using the goal–management component of ELWMS.KOM was provided and the evaluation and control groups were both using a similar version of ELWMS.KOM in order to make them more comparable to the experimental groups.

Additionally, following research questions were of interest:

- What are the differences between learners that organize their found web resources with folders (the control group) and learners that set goals prior to learning (the treatment groups)?
- What are the differences between learners that are explicitly instructed to execute metacognitive processes (the control group and the first treatment group getting indirect scaffolds) and learners

[8] Unfortunately, the logged data did not include which properties of the resource were edited. This was fixed in the 2^{nd} version of ELWMS.KOM.

that are free to use the functionality to support their metacognitive processes (the treatment group prompted by direct scaffolds)? Thus, what are the benefits of providing direct scaffolds?

Study Design

104 students could be won for participating in this study. 74.5% are students of Psychology and 13.2% students of Education between 19 and 28 years of age with 90.2% in their first two semesters. Due to the field of study, the majority of participants were women (72.6%) and 88.7% of the participants speak German as first language. There was no overlap between the participants of the exploratory study and this study. The participants were randomly allocated to three groups:

- The Control Group (CG_2, $n = 34$) was using ELWMS.KOM with a stripped–down goal–management component that did not exhibit the goal–setting functionality. *Goals* were named *Folders* in order not to bias the participants, and goals could not be activated or attributed progress. Still, participants of CG_2 were able to put resources and snippets thereof in a folder and access the different displays of the collected data.
- The first Treatment Group (TG_21, $n = 35$) used the goal–management component with the complete functionality but was not given instructions on how to organize their search process. Hence, this group realized indirect scaffolds as given in section 6.2.
- The second Treatment Group (TG_22, $n = 35$) used the same tool with integrated metacognitive prompts aimed at activating and supporting the metacognitive processes *defining relevant goals*, *keeping the active goal in mind*, *finding relevant pages*, *importing relevant information*, *assigning relevant information to the relevant goal* and *learning relevant information*. For example, before beginning the search process (i.e. in the pre–action phase, cf. section 6.2.1), the learners were instructed to set goals for their search. Further, during search, instructions to reflect on whether the found information was relevant for the currently followed goal were given (see figure 6.6). Like in the exploratory study, five minutes before the end of the learning phase, this group was instructed to reflect on their results.

The overall study was performed in two sessions on two different days for each participant (see figure 6.7). The first session was exclusively for training with the respective tool variant and the second was the search task. The first session was always held the day before the search task and gave the participants the possibility to get to know the handling of the respective tool variant they would use on the search task. First, they watched an introductory presentation in the respective version of ELWMS.KOM, showing common tasks and the functionality of the tool. Then, the participants were given a small search task in a topic they were confident with, where they could apply the functionality of their tool variant. Further, demographic data and data about the participants' self–conceptions about their computer skills (e.g. estimation of their familiarity in using computers and knowledge about relevant computer– and internet–related concepts) and skills of self–regulated web search (i.e. the competencies to plan and structure their learning processes, based on items of a standardized questionnaire according to PANAS [111] like in the first study) were collected.

The second session on the day 2 (cf. figure 6.7) was designed to be approximately 1.5 hours in length. Participants were given a first achievement test (multiple choice) about the *Classical Antiquity*. 14 multiple–choice test questions were formulated that could be answered by information contained in different Wikipedia articles. As some articles cover a lot of information, the minimum number of resources that contain all information needed for answering the questions was six. After each question the

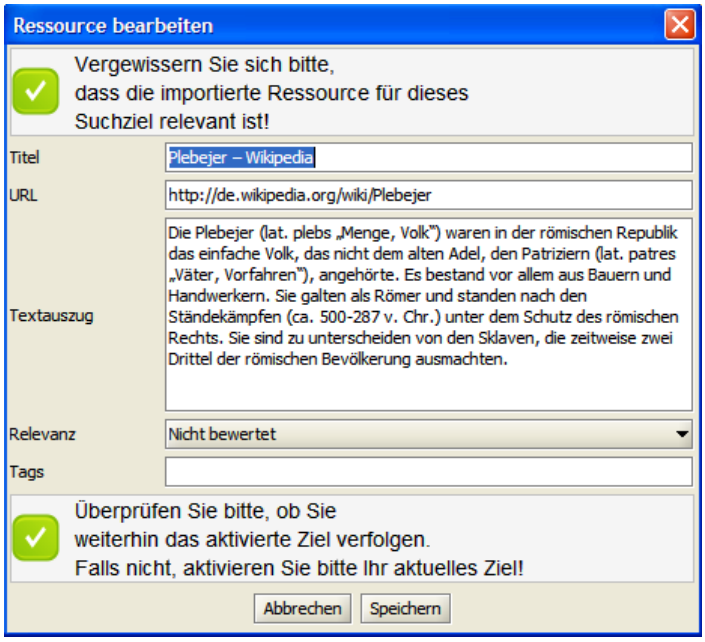


Figure 6.6: Screenshot of ELWMS.KOM's "Add Resource" dialog with example of a metacognitive prompt, requesting the learner to reflect whether the imported web resource is relevant for the current research goal. The lower message prompts the user to check if she is still following the currently activated goal.

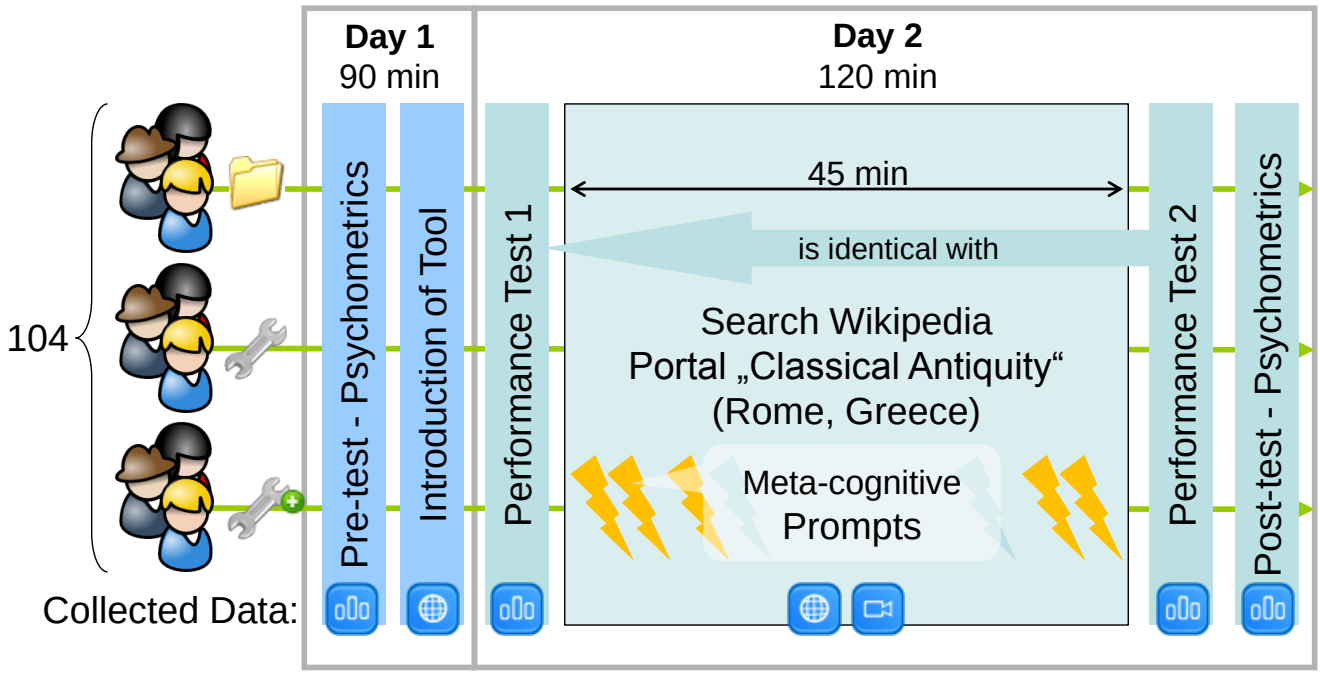


Figure 6.7: The design of the second session of the 2^{nd} Study. The icons at the bottom of each respective phase denote the kind of data obtained (questionnaire data in Pre–, Post– and Performance test, log data in the introduction and evaluation and screen recordings only in the evaluation itself).

participants were asked to state how certain they were with answering this question (from the extremes *I guessed* to *I know and I am sure* in four steps). There were ten different versions of the test, which differed in the order the questions were provided. Participants were given the hint that they would receive exactly the same test again after the learning episode. Each participant received a feedback on her individual test performance. Ten questions which were either answered incorrectly or with uncertainty were provided for the first five minutes of the learning episode. This enabled competent learners to identify knowledge gaps in the achievement test and to re–formulate these into search goals in order to finally answer them correctly. During the search process, participants were given updates about the time left after 20 minutes and five minutes before the end. After learning, the achievement test was administered to the participants a second time. Finally, the participants were asked to answer some questions about their learning and their experiences during the web search, their emotions according to PANAS, their usage of the goal–management component and its functionality. At the beginning and the end of the learning episode, the current levels of motivation and self–efficacy were gathered in a survey. Further, like in the exploratory study, questionnaire and log file data were stored and screen–captures of the search process were taken.

Selected Results of the Study

As in the exploratory study, the topic *Classical Antiquity* was chosen for the learning materials. In order to estimate their prior knowledge in this topic, the students were asked to state how much they knew about the Roman Antiquity (83% stated they have only "rather marginal" or "little" background, whereas only 2% said to have a "very good" knowledge about this subject) and Greek Antiquity (where only 1% of the participants claimed to have a "very good" knowledge about, in contrast to 86% of the participants stated to have a "rather marginal" or "little" background).

Due to the topic–relatedness of given tasks, goals were usually set in a topic–oriented way, process–oriented goals (e.g. "I need to get an overview of . . . ") were rarely set. The results presented below are all based on the log files and the questionnaires.

Selected Group Differences

In order to analyse the differences between all three groups including differences within specific phases of action, one way ANOVAs with quantitative log data as the independent variables were conducted. Table 6.4 presents selected significant results. In contrast to the first, the group differences were calculated for each of the learning phases.

These results show that, as presumed, in the pre–action phase the three groups differ in terms of numbers of goals/folders created and edited, links followed, as well as number of imported, viewed and edited resources. Further, the number of viewed resources and links followed in the post–action phase varied between groups. A difference between groups over all phases was encountered for moved goals/folders. These results in general indicate different approaches of web search for learners of different groups. Some learners seem to have searched in a very structured manner by first defining their search goals instead of immediately starting to browse and persist resources. These learners also seem to have reduced distracting activities like aimless browsing at the end of the learning phase in order to prepare for the post–test. For example, in figure 6.8, a timeline of actions of three selected participants from the Treatment Groups is displayed. Participant IZ42 of $TG_2 1$ who was not given prompts did not structure her approach properly, the execution of goal–setting and resource attachment process show that she performed the learning task ad–hoc, creating goals and adding resources when needed. In contrast, participants 8IRH and PENT (who were part of $TG_2 2$) show to have followed an explicit goal–setting phase

Variable	Phase of Action	ANOVA
Creation of Goal/Folder	Pre	$F(2,102) = 7.729$, $p < .01^{**}$
Editing Goals/Folder	Pre	$F(2,102) = 3.801$, $p < .05^{*}$
Moving Goals	All	$F(2,102) = 3.600$, $p < .05^{*}$
Following new Link	Pre	$F(2,102) = 6.280$, $p < .01^{**}$
	Post	$F(2,102) = 6.885$, $p < .01^{**}$
Import Resource	Pre	$F(2,102) = 5.106$, $p < .01^{**}$
View Resource	Post	$F(2,102) = 3.827$, $p < .05^{*}$
Editing Resource	Pre	$F(2,102) = 3.105$, $p < .05^{*}$

Table 6.4: 2^{nd} Study: Selected group differences of CG_2 versus $TG_2 1{+}2$ based on participants' actions during the respective phases obtained with the log files. F =F–value, p =niveau of significance. * denotes significance ($p < .05$), ** denotes strong significance ($p < .01$).

before starting to browse and add the resources. However, the proceeding of participant 8IRH shows that she did not have an explicit reflection phase (e.g. she never re–opened the stored resources again), whereas participant PENT took a few minutes at the end in order to reflect on her findings. According to the theory of SRL, this is an important part of the learning process. Although this did not have a direct effect on learning performance, this shows that the structure of the learning episode can be derived from the logfiles.

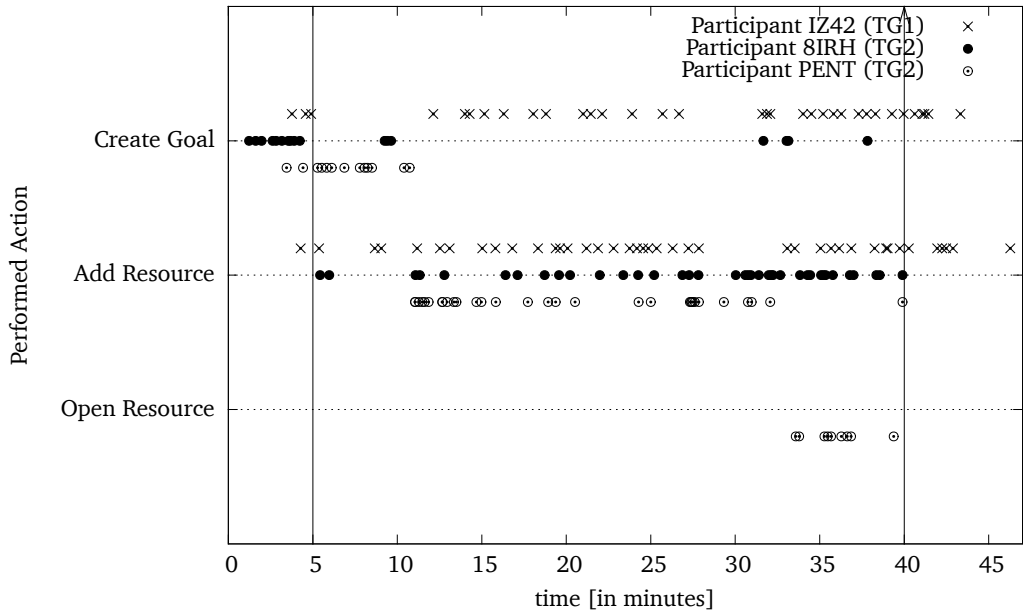


Figure 6.8: 2^{nd} Study: Timeline of selected user actions of three participants. The vertical lines denote the separation between the phases, i.e. the time instructions were given to proceed to the next phase.

As these results show only the *presence* of significant group differences, specific differences between the respective groups defined in the research questions were further inspected and contrasted.

In order to analyse the research question what the differences are between learners that organize their found web resources using the folder metaphor and learners that set goals prior to learning, the two treatment groups $TG_2 1$ and $TG_2 2$ that were provided with the goal setting function and the control group CG_2 that applied folders were contrasted. Table 6.5 shows the results of the evaluation of Control

Action	Groups	Phase	Groups' Mean values	Student's t–test
Opened Web Resources	$TG_21 \cap TG_22 > CG_2$	Pre	$M_{CG_2} = 3.79, M_{TG_21} = 3.49, M_{TG2_2} = 1.39$	$t(102) = 2.018, p < .05^{*}$
	$TG_22 < CG_2 \cap TG_21$	Pre	$M_{CG_2} = 3.79, M_{TG_21} = 3.49, M_{TG2_2} = 1.39$	$t(102) = -3.866, p < .01^{**}$
	$TG_21 \cap TG_22 < CG_2$	Post	$M_{CG_2} = 1.12, M_{TG_21} = 0.60, M_{TG2_2} = 0.33$	$t(102) = 2.887, p < .01^{**}$
	$TG_22 < CG_2 \cap TG_21$	Post	$M_{CG_2} = 1.12, M_{TG_21} = 0.60, M_{TG2_2} = 0.33$	$t(102) = -3.415, p < .01^{**}$
Create Goals	$TG_21 \cap TG_22 > CG_2$	Pre	$M_{CG_2} = 5.47, M_{TG_21} = 5.91, M_{TG2_2} = 8.92$	$t(102) = -2.068, p < .05^{*}$
	$TG_22 > CG_2 \cap TG_21$	Pre	$M_{CG_2} = 5.47, M_{TG_21} = 5.91, M_{TG2_2} = 8.92$	$t(102) = 3.020, p < .01^{**}$
	$TG_22 < CG_2 \cap TG_21$	Action	$M_{CG_2} = 2.56, M_{TG_21} = 2.97, M_{TG2_2} = 1.69$	$t(102) = 4,296, p < .01^{**}$
Edit Goals	$TG_21 \cap CG_2 < TG_22$	Action	$M_{CG_2} = 0.24, M_{TG_21} = 0.23, M_{TG2_2} = 0.69$	$t(102) = -2.768, p < .01^{**}$
Restructure Goals	$TG_22 > CG_2 \cap TG_21$	All	$M_{CG_2} = 2.18, M_{TG_21} = 1.74, M_{TG2_2} = 4.14$	$t(102) = 2.253, p < .05^{*}$
	$TG_21 \cap TG_22 < CG_2$	Pre	$M_{CG_2} = 0.47, M_{TG_21} = 0.91, M_{TG2_2} = 2.19$	$t(102) = -2.783, p < .01^{**}$
Revisit Resources	$TG_21 \cap TG_22 > CG_2$	Post	$M_{CG_2} = 0.68, M_{TG_21} = 0.83, M_{TG2_2} = 2.42$	$t(102) = -1.964, p < .05^{*}$
	$TG_22 > CG_2 \cap TG_21$	Post	$M_{CG_2} = 0.68, M_{TG_21} = 0.83, M_{TG2_2} = 2.42$	$t(102) = 2.200, p < .05^{*}$
Search operation	$TG_21 > TG_22 \cap CG_2$	Action	$M_{CG_2} = 3.62, M_{TG_21} = 4.69, M_{TG2_2} = 5.58$	$t(102) = -1.790, p < .05^{*}$
Goal Activation	$TG_22 > TG_21$	All	$M_{TG_21} = 1.26, M_{TG2_2} = 5.47$	$t(69) = 3,463, p < .01^{**}$

Table 6.5: 2^{nd} Study: Results of Evaluation of compound Control and Treatment Group Differences. * denotes significance ($p < .05$), ** denotes strong significance ($p < .01$). All variables are obtained with log files.

and Treatment Group differences in the respective phases. The significance levels are computed by using a 1–tailed, independent two–sample t–test with unequal sample sizes and variance. TG_21 and TG_22 significantly set more goals, specifically in the first phase before learning and opened less new web pages in the browser during the pre–action and post–action phase spending more time with the processes of planning and reflection. This means that they first organized their course of actions before starting to learn. Additionally, they restructured their goal hierarchy more often while planning, which is an indication to be the result of a detailed planning process. Further, the treatment groups updated their goals and performed more searches in Wikipedia during the action phase more often than the control group, showing that they monitored their progress and based on the current state altered the information they had already searched for. This may be due to a more goal–oriented approach, identifying and re–evaluating knowledge gaps and acting on those new or changed information needs. Finally, the treatment groups more often revisited the collected relevant resources after learning, distilling the relevant information and memorizing it for the post–test.

To analyse the research question whether there are differences between learners that are explicitly instructed to execute metacognitive processes and learners that do not receive metacognitive prompts, TG_22, which had received direct support during learning were contrasted with TG_21 and CG_2, which were only indirectly supported. TG_22 set more goals, especially in the pre–action phase, whereas later they actually set less goals, meaning they took more time to plan their course of action, approaching the search task in a more goal–directed way and performing the search process more efficient. Further, TG_22 opened less web resources while browsing, having previously identified their knowledge gaps and looking specifically for relevant resources. Participants in TG_22 were more often reorganizing their goals, regulating the current state and opened significantly less new pages before and after learning, meaning they acted more efficiently and kept closer to their set goal. Further, after having learned, they more often reflected on found relevant resources. Participants using ELWMS.KOM with metacognitive prompts (TG_22) used the goal activation functionality far more frequently than the group without prompts. This means that learners in TG_22 significantly monitored their progress more often than TG_21.

In conclusion, these results show that using ELWMS.KOM for setting goals affects the way learners approach their search process using web resources: they execute more metacognitive processes, plan in

a more detailed way, monitor their progress better and react on changed circumstances and more often reflect on their learning outcomes and found web resources.

Similarly to the exploratory study presented in section 6.4.2, there were no significant differences in terms of performance (i.e. more correctly answered questions) in a group comparison. This might be due to the short scope of this study and that third variables (e.g. certainty when answering questions or the relevance of found resources) were not included in this study.

Selected Correlations

To investigate further dependencies, several correlations between variables accounting for different patterns within different groups were calculated. A selection of significant correlations is presented in table 6.6.

Group	Variable 1	Variable 2	Correlation r
CG_2	Computer Literacy$^{\beta}$	Perceived Benefits of ELWMS.KOM$^{\beta}$	.364*
CG_2	Computer Literacy$^{\beta}$	Would use ELWMS.KOM$^{\beta}$	.445**
CG_2	Computer Literacy$^{\beta}$	Perceived Benefits of Snippets$^{\beta}$	.472**
All	Computer Literacy$^{\beta}$	Number of Goals$^{\alpha}$	.356**
TG_21; TG_22	Search Literacy$^{\beta}$	Number of Goals$^{\alpha}$	.292*; .304*
CG_2; TG_22	PANAS "active"$^{\beta}$	Number of Goals$^{\alpha}$	.325*; $-$.331*
All	PANAS positive$^{\beta}$	Opened Web Resources$^{\alpha}$	$-$.256**
TG_22	PANAS negative$^{\beta}$	Opened Web Resources$^{\alpha}$	.436**

Table 6.6: 2^{nd} Study: Selected significant correlations between variables, with significance $^{*}: p < .05$, $^{**}: p < .01$. Variables marked with $^{\alpha}$ designate data gained from logfiles, variables marked with $^{\beta}$ were collected in questionnaires.

In the Control Group CG_2, the higher the participant's computer literacy was rated by herself, the more she thought TEL using web resources benefits from using ELWMS.KOM, the better she liked the goal–management extension to ELWMS.KOM and the more valuable she estimated the snippet functionality for TEL. In both treatment groups, computer literacy was not correlated to those variables. This might indicate that participants of the CG implicitly knew how to use the stripped–down version of ELWMS.KOM if they had a high computer literacy. Participants of the other groups, however, were supported in setting goals, monitoring them and reflecting on the learning process. Therefore, giving them that much support might have neutralized the influence of computer literacy on organizing their search process.

Further, creation of goals correlated with computer literacy in all groups, meaning participants describing themselves as competent in using computers set more goals. Moreover, participants of the treatment groups set more goals the more confident they were of their ability to perform an efficient web search. Curiously, there were clear correlations between the emotion to be "active" and the amount of goals/folders created — for CG_2, it was positive, meaning that participants in this group felt themselves to be more active when setting more goals, whereas for the TG_22 it was negative —- the more goals a participant of this group set, the less active she felt. This might indicate that a strong direct support, among all the positive impact, might cause learners to feel less active. To be provided with more freedom, however, might cause the feeling of activeness in terms of being in charge of ones' own actions.

Eventually, the more web resources were opened, the less positive emotions the participants in all groups had and the less activated the participants felt. Additionally, for TG_22, negative emotions (PANAS) were correlated to the number of opened resources. This means that browsing the web re-

sources for information aimlessly (thus browsing a lot of different web resources, eventually becoming *lost in hyperspace*) affects the emotions of learners negatively. Still, participants in the Control Group did not have negative emotions when browsing more pages. This might be due to the fact that learners who did not set search goals did less encounter their browsing of many resources as being ineffective and accordingly experienced less negative emotions.

6.4.4 Conclusions of Both Studies

The two presented studies aimed at representing a typical, proximal information search using web resources in order to meet a specific information need. As a search in the whole Web is infeasible for a study that aims to provide comparable results, Wikipedia was used as a closed subset that is representative for the Web's. Both studies indicate the benefits of supporting meta–cognitive, self–regulated learning processes. The results show that using ELWMS.KOM for setting goals affects the way learners approach learning processes using web resources: they execute more meta–cognitive processes, plan in a more–detailed way, monitor their progress better and react on changed circumstances and more often reflect on their learning outcomes and found web resources. Both studies did not yield significant performance differences between the groups, however, this was expected due to the short duration of the study session (cf. section 6.4.2).

The implemented functionality was well received by the participants of both studies: most of them (87.2%) saw the need to being able to store only small, relevant snippets of a web resource in learning with web resources. As common web browsers do not provide this functionality, whereas ELWMS.KOM offers this advantage. Further, 77.0% of all participants stated they would like to use ELWMS.KOM or a comparable goal–setting tool for their learning episodes. This shows that learners indeed see the need to organize their learning processes and expect they could benefit from such a tool.

6.5 Conclusions and Further Steps

In this chapter, a goal–management component for ELWMS.KOM has been presented that is based on theoretical principles of SRL and the term *scaffolds* was introduced for denoting functionality supporting metacognitive processes during learning episodes. The goal–management component has been evaluated in two studies. Results show that using the tool for setting goals affects the way learners approach searching using web resources: they execute more metacognitive processes, plan in a more detailed way, monitor their progress better and react on changed circumstances and more often reflect on their learning outcomes and found web resources. However, significant differences with regard to learning performance between the groups could not be found. This is not surprising, as the duration of the examined learning episodes was only 45 minutes and the benefits of goal–setting emerge not until the need of a detailed goal structure arises. Yet, the promising results indicate that learners being enabled to regulate their metacognitive processes benefit from this support. The presented studies caused the inclusion of the *Goal* tag type in ELWMS.KOM as a valuable addition to assist learners in structuring and regulating their learning.

Due to the complexity of the evaluation design, only personal learning settings were targeted. A very interesting research question is how learners fare in community–based, collaborative learning settings. The *Crokodil* project [160], a successor to ELWMS.KOM, is currently pursuing this field of research. It is expected that also here, goal–setting is beneficial for learners. Particularly, other learners are expected to profit from personal goal hierarchies by being able to consume a pre–organized collection of learning

resources. This is expected to foster the collaboration and cooperation between learners that have a related information need and will be investigated in further studies.

7 Conclusions and Further Work

This concluding chapter presents a summary of the contributions of this thesis. Further, an outlook of the state of ELWMS.KOM is given and future work is identified that has been established by this thesis.

7.1 Summary of Contributions

In this thesis, the notion of RBL was presented and the applicability of the concept of Learning Objects to RBL was explored. The conceptual design and implementation of the overall platform of ELWMS.KOM has been developed as a foundation for the support of RBL. Considerations including the nature and organization of learning materials used for RBL, didactic implications of self–direction and the role of the learner have been incorporated into the design of ELWMS.KOM.

Building on an analysis of user data in ELWMS.KOM, it has been shown that LRs in RBL are usually rather short and are often composed in different languages. Based on these properties, the requirements for a content–based recommendation approach have been derived. In related work, Explicit Semantic Analysis has been identified as an adequate approach that allows to infer over semantic relatedness between texts. This thesis systematically explored and evaluated the impact of concept and term reduction in ESA, reducing the overall computational complexity of the approach. Further, the applicability of ESA on cross–language setting was examined, providing a novel concept mapping approach and evaluating its performance. Eventually, ESA was enhanced by novel extensions to ESA that incorporate further semantic characteristics of Wikipedia. This approach named XESA was tested on a novel semantic corpus and showed to surpass ESA's quality on document comparisons by 7%.

Further, an algorithm to automatically segment web resources into coherent fragments that enables usability support in ELWMS.KOM was presented. Based on an analysis of related work and its short–comings, a novel approach to coherently segment web resources was presented, taking into account re–occurring structural patterns. An evaluation design was derived and in a user study the presented approach was evaluated, showing to yield good results.

A novel approach to web genre detection called LIGD that is language–independent and targets at supporting ELWMS.KOM to automatically classify the web genre of a resource in order to provide metadata was introduced. In an analysis of a social bookmarking application it was shown that the targeted web genres belong to the most–used tags. Therefore, existing features were reviewed and novel features were presented that capture the pattern structure of a web resource. As the proposed approach exclusively takes into account structural features of the web resource, it proved to be language–independent. A corpus of multilingual instances of the targeted web genres was built and several evaluations were performed that support the applicability of this approach with a noteworthy accuracy of up to 96.2%.

Besides the technological explorations, an implementation of the notion of scaffolds in Self–Regulated Learning in ELWMS.KOM was introduced and its impact on the application of metacognitive processes was stated. Two user studies were presented that substantiate the theoretical benefits of supporting a goal–directed advancement in self–directed RBL.

7.2 Future Perspectives

There are several possibilities to enhance ELWMS.KOM. A topic that has only be touched in this thesis is the support of collaborative and cooperative learning settings. With social learning, the exchange of

information can be leveraged by supportive functionality that allows learners to disseminate, share, pool and organize their collected information in order to build a common knowledge base. Further, features in ELWMS.KOM that allow synchronous or asynchronous communication between learners is not yet supported. These challenges are targeted in the *Crokodil* project [160], a successor to ELWMS.KOM. *Crokodil* is currently pursuing this field of research, employing many features of ELWMS.KOM with the goal of supporting collaborative RBL.

Of the separate technological contributions of this thesis, especially XESA is interesting for a further investigation, as it is applicable in a diversity of usage scenarios, including Information Retrieval scenarios and general recommendation algorithms. A research in progress aims to enhance XESA by taking the relation weights between articles and categories into account. This means that a measure of relation between linked articles or categories is weighted according to the semantic strength between these concepts. This approach is expected to perform better than XESA that only takes into account the existence of links, but does not add weights to this relation. Further, the cross–lingual mapping technique using Meta Cross–Language Links should benefit considerably from the same weighting. Considering different applications, the calculation of semantic relatedness provides a stable measure and performs well, making it interesting to follow and evaluate other usage scenarios. An issue of XESA is still its computation time, so an important improvement of its applicability would be to reduce its computational complexity to a greater extent. Thus, the research on XESA should be pursued furthermore and it should be specifically used as a basis for providing recommendations.

Bibliography

[1] J. Allsopp. Semantics in the wild. `http://westciv.typepad.com/dog_or_higher/2005/11/real_world_sema.html`, retrieved 2010-11-05, Nov 2005.

[2] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. In *Proceedings of ACM Hypertext 2003*, volume 4, pages 38–47. ACM, 2003.

[3] M. Anderka and B. Stein. The ESA retrieval model revisited. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 670–671. ACM, 2009.

[4] G. Attwell. Personal Learning Environments — The Future of eLearning? *eLearning Papers*, 2(1):1–8, Jan 2007.

[5] R. Azevedo. Using Hypermedia as a Metacognitive Tool for Enhancing Student Learning? The Role of Self-Regulated Learning. *Educational Psychologist*, 40(4):199–209, 2005.

[6] R. Azevedo, J. G. Cromley, and D. Seibert. Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology*, 29(3):344–370, 2004.

[7] R. Azevedo and A. F. Hadwin. Scaffolding self–regulated learning and metacognition — Implications for the design of computer–based scaffolds. *Instructional Science*, 33(5):367–379, 2005.

[8] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[9] S. Baluja. Browsing on small screens: recasting web–page segmentation into an efficient machine learning framework. In *Proceedings of the 15th international conference on World Wide Web*, pages 33–42. ACM Press New York, NY, USA, 2006.

[10] A. Bandura. Social Cognitive Theory of Self-Regulation. *Organizational Behavior and Human Decision Processes*, 50:248–287, 1991.

[11] A. Bandura and D. H. Schunk. Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, 41(3):586–598, 1981.

[12] M. Bannert. Effekte metakognitiver Lernhilfen auf den Wissenserwerb in vernetzten Lernumgebungen. *Zeitschrift für Pädagogische Psychologie*, 17(1):13–25, 2003.

[13] M. Bannert, M. Hildebrand, and C. Mengelkamp. Effects of a metacognitive support device in learning environments. *Computers in Human Behavior*, 25(4):829–835, 2009.

[14] Z. Bar-Yossef and S. Rajagopalan. Template detection via data mining and its applications. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 580–591, New York, NY, USA, 2002. ACM.

[15] C. Barrit, D. Lewis, and W. Wieseler. Cisco Systems Reusable Information Object Strategy — Definition, Creation Overview, and Guidelines. Whitepaper, Cisco Systems, Inc., Jun 1999. v. 3.0.

[16] M. Bauer, R. Maier, and S. Thalmann. Metadata Generation for Learning Objects: An Experimental Comparison of Automatic and Collaborative Solutions. In M. H. Breitner, F. Lehner, J. Staff, and U. Winand, editors, *E–Learning 2010*, pages 181–195. Physica–Verlag Heidelberg, 2010.

[17] L. M. Baumgartner. *Adult Learning Theory*, chapter Self-Directed Learning: A Goal, Process, and Personal Attribute, pages 23–28. Number 392 in Information Series. Center on Education and Training for Employment, Columbus, Ohio, USA, 2003.

[18] P. Baumgartner. *Campus 2004 — Kommen die digitalen Medien an den Hochschulen in die Jahre?*, chapter Didaktik und Reusable Learning Objects (RLOs), pages 311–327. Waxmann, Münster, 2004.

[19] B. F. Benz. *Improving the Quality of E-Learning by Enhancing Self-Regulated Learning. A Synthesis of Research on Self-Regulated Learning and an Implementation of a Scaffolding Concept.* PhD thesis, Technische Universität Darmstadt, 2010.

[20] B. F. Benz, S. Polushkina, B. Schmitz, and R. Bruder. Developing Learning Software for the Self–Regulated Learning of Mathematics. In M. B. Nunes. and M. McPherson, editors, *IADIS Multi Conference on Computer Science and Information Systems*, pages 200–204. IADIS Press, 2007.

[21] B. F. Benz and B. Schmitz. Fostering self–regulated web search: Evaluating a goal management tool. In *Proceedings of the European Association for Research on Learning and Instruction (EARLI) Metacognition SIG*, Ioannina, Greece, May 2008.

[22] K. Bischoff, E. Herder, and W. Nejdl. Workplace Learning: How We Keep Track of Relevant Information. In E. Duval, R. Klamma, and M. Wolpers, editors, *Proceedings of EC–TEL 2007*, volume 4753 of *LNCS*, pages 438–443. Springer Verlag Berlin Heidelberg, 2007.

[23] G. Biswas, K. Leelawong, K. Belynne, and B. Adebiyi. Case Studies in Learning by Teaching Behavioral Differences in Directed versus Guided Learning. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 828–833, Stresa, Italy, 2005.

[24] G. Biswas, K. Leelawong, K. Belynne, K. Viswanath, D. Schwartz, and J. Davis. Developing Learning by Teaching Environments That Support Self–Regulated Learning. In *Intelligent Tutoring Systems 2004*, number 3320 in Lecture Notes in Computer Science, pages 730–740. Springer, 2004.

[25] M. Boekarts and A. Minnaert. Self-regulation with respect to informal learning. *International Journal of Educational Research*, 31:533–544, 1999.

[26] E. S. Boese. Stereotyping the Web: Genre Classification of Web Documents. Master's thesis, Colorado State University, Mar 2005.

[27] E. S. Boese and A. E. Howe. Effects of Web Document Evolution on Genre Classification. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 632–639, New York, NY, USA, 2005. ACM.

[28] D. Böhnstedt. *Semantisches Tagging zur Verwaltung von webbasierten Lernressourcen — Modelle, Methoden und eine Plattform zur Unterstützung Ressourcen–basierten Lernens*. PhD thesis, Technische Universität Darmstadt, Apr 2011. to appear in 2011.

[29] D. Böhnstedt, P. Scholl, C. Rensing, and R. Steinmetz. Collaborative Semantic Tagging of Web Resources on the Basis of Individual Knowledge Networks. In G.-J. Houben, G. McCalla, F. Pianesi, and M. Zancanaro, editors, *Proceedings of First and Seventeenth International Conference on User Modeling, Adaptation, and Personalization UMAP 2009*, volume Lecture Notes in Computer Science, pages 379–384. Springer-Verlag Berlin Heidelberg, Jun 2009.

[30] D. Böhnstedt, P. Scholl, C. Rensing, and R. Steinmetz. Modeling Personal Knowledge Networks to Support Resource Based Learning. In K. Tochtermann and H. Maurer, editors, *Proceedings of 9th International Conference on Knowledge Management and Knowledge Technologies (I-KNOW'09)*, pages 309–316. Verlag der Technischen Universität Graz, Austria, Universiti Malaysia Sarawak, Malaysia, and Know-Center, Austria, Sep 2009.

[31] D. Böhnstedt, P. Scholl, C. Rensing, and R. Steinmetz. Enhancing an Environment for Knowledge Acquisition based on Web Resources by Automatic Tag Type Identification. In M. E. Auer and J. Schreurs, editors, *Proceedings of International Conference on Computer-aided Learning 2010 (ICL 2010)*, pages 380–389, Kassel, Sep 2010. Kassel University Press.

[32] B. Bornemann-Jeske. Barrierefreies Webdesign zwischen Webstandards und Universellem Design. *Information Wissenschaft und Praxis*, 56(8):418ff, 2005.

[33] J. Bos and M. Nissim. Cross–Lingual Question Answering by Answer Translation. In *Workshop of Cross-Language Evaluation Forum (CLEF)*, 2006.

[34] C. Bouras and A. Konidaris. Web Components: A Concept for Improving Personalization and Reducing User Perceived Latency on the World Wide Web. In *Proceedings of the 2nd International Conference on Internet Computing (IC2001)*, volume 2, pages 238–244, Las Vegas, Nevada, USA, Jun 2001.

[35] T. Boyle. Design Principles for Authoring Dynamic, Reusable Learning Objects. In A. Williamson, C. Gunn, A. Young, and T. Clear, editors, *Winds of Change in the Sea of Learning: Proceedings of the 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*, pages 57–64. Institute of Technology Auckland, 2002.

[36] S. Braun, A. Schmidt, A. Walter, and V. Zacharias. Von Tags zu semantischen Beziehungen: kollaborative Ontologiereifung. In B. Gaiser, T. Hampel, and S. Panke, editors, *Good Tags — Bad Tags: Social Tagging in der Wissensorganisation*, volume 47 of *Medien in der Wissenschaft*, chapter 4, pages 163–173. Waxmann, 2008.

[37] P. S. Breivik. Information Literacy. *Bulletin od the Medical Library Association*, 79(2):226–229, 1991.

[38] C. Brooks, S. Bateman, W. Liu, G. McCalla, J. Greer, D. Gasevic, T. Eap, G. Richards, K. Hammouda, S. Shehata, M. Kamel, F. Karray, , and J. Jovanović. Issues and Directions with Educational Metadata. In *Proceedings of the Third Annual International Scientific Conference of the Learning Object Repository Research Network*, Montreal, Nov 2006.

[39] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM.

[40] I. Buchem and H. Hamelmann. Microlearning a strategy for ongoing professional development. *eLearning Papers*, 1(21):1–15, Sep 2010.

[41] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[42] D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block–level link analysis. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 440–447, New York, NY, USA, 2004. ACM.

[43] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi–relational networks. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 445–452. Springer, 2005.

[44] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. VIPS: a vision–based page segmentation algorithm. Technical Report MSR-TR-2003-79, Microsoft Research, Nov 2003.

[45] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Block–based Web Search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 456–463, New York, NY, USA, 2004. ACM.

[46] S. Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 211–220, New York, NY, USA, 2001. ACM.

[47] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.

[48] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 208–216, New York, NY, USA, 2001. ACM.

[49] A. J. Champandard. The easy way to extract useful text from arbitrary html. `http://ai-depot.com/articles/the-easy-way-to-extract-useful-text-from-arbitrary-html/`, retrieved 2010-11-05, Apr 2007.

[50] D. Chandler. An Introduction to Genre Theory. `http://www.aber.ac.uk/media/Documents/intgenre/`, retrieved 2010-05-14, 2000.

[51] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien. Translating Unknown Queries with Web Corpora for Cross–Language Information Retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 146–153. ACM, 2004.

[52] P.-A. Chirita, S. Costache, S. Handschuh, and W. Nejdl. P-TAG: Large Scale Automatic Generation of Personalized Annotation TAGs for the Web. In *Proceedings of the WWW Conference 2007*, pages 845–854. IW3C2, 2007.

[53] B. Christos, K. Vaggelis, and M. Ioannis. Web Page Fragmentation for Personalized Portal Construction. In *ITCC '04: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04)*, volume 2, page 332, Washington, DC, USA, 2004. IEEE Computer Society.

[54] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab. Explicit vs. Latent Concept Models for Cross–Language Information Retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'09)*, pages 1513–1518, 2009.

[55] J. Conklin. Hypertext: A survey and introduction. *IEEE Computer*, 20(9):17–41, 1987.

[56] I. Cramer. How Well Do Semantic Relatedness Measures Perform? A Meta-Study. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 59–70, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[57] S. Debnath, P. Mitra, and C. L. Giles. Automatic extraction of informative blocks from webpages. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 1722–1726, New York, NY, USA, 2005. ACM Press.

[58] S. Debnath, P. Mitra, N. Pal, and C. L. Giles. Automatic Identification of Informative Sections of Web Pages. *IEEE Transactions on Knowledge and Data Engineering*, 17(9):1233–1246, 2005.

[59] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[60] N. Dewdney, C. VanEss-Dykema, and R. MacMillan. The Form is the Substance: Classification of Genres in Text. In *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*, pages 37–45, 2001.

[61] R. Domínguez García, A. Berlea, P. Scholl, D. Böhnstedt, C. Rensing, and R. Steinmetz. Improving Topic Exploration in the Blogosphere by Detecting relevant Segments. In *Journal of Universal Computer Science: Proceedings of the I-Know 2009*, pages 177–188. Verlag der Technischen Universität Graz, 2009.

[62] S. Downes. Learning Objects: Resources For Distance Education Worldwide. *International Review of Research in Open and Distance Learning*, 2(1):1–35, 2001.

[63] S. Downes. E-learning 2.0. *eLearn magazine*, 2005(10):1–9, Oct 2005.

[64] P. F. Drucker. *Management: Tasks, Responsibilities, Practices*. Harper & Row Publishers, New York, 1974.

[65] E. Elgersma and M. de Rijke. Learning to Recognize Blogs: A Preliminary Exploration. In *EACL 2006 Workshop on New Text — Wikis and Blogs and Other Dynamic Text Sources*, pages 24–31, 2006.

[66] A. Faatz. *Ein Verfahren zur Anreicherung fachgebietsspezifischer Ontologien durch Begriffsvorschläge*. PhD thesis, Technische Universität Darmstadt, Nov 2004.

[67] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. MIT Press, 1998.

[68] S. Ferrández, A. Toral, O. Ferrández, A. Ferrández, and R. Munoz. Applying Wikipedia's Multilingual Knowledge to Cross–Lingual Question Answering. *Natural Language Processing and Information Systems*, 4592:352–363, 2007.

[69] A. P. Field. *Discovering statistics using SPSS*. SAGE publications, 2009.

[70] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems (TOIS)*, 20(1):116–131, Jan 2002.

[71] A. Finn and N. Kushmerick. Learning to Classify Documents According to Genre. In *IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[72] A. Finn and N. Kushmerick. Learning to Classify Documents According to Genre. *Journal of the American Society for Information Science and Technology (JASIST)*, 57(11):99–113, Jul 2006.

[73] R. Flesch. A new readability yardstick. *Journal Of Applied Psychology*, 32:221–233, 1948.

[74] A. Föhr. Extraction of Structurally Coherent Segments from Web Pages Using a Hybrid Recursive Approach. Diploma thesis, Technische Universität Darmstadt, Darmstadt, Germany, Aug 2008.

[75] A. Foster and N. Ford. Serendipity and information seeking: an empirical study. *Journal of Documentation*, 59(3):321–340, 2003.

[76] H. F. Friedrich and H. Mandl. *Lern– und Denkstrategien. Analyse und Intervention*, chapter Lern– und Denkstrategien — ein Problemaufriß, pages 3–54. Hogrefe, 1992.

[77] E. Gabrilovich. *Feature Generation for textual Information Retrieval using World Knowledge*. PhD thesis, Israel Institute of Technology, 2006.

[78] E. Gabrilovich and S. Markovitch. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *Proceedings of the Twenty–First National Conference on Artificial Intelligence*, pages 1301–1306. American Association for Artificial Intelligence, AAAI Press, 2006.

[79] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.

[80] D. Gibson, K. Punera, and A. Tomkins. The volume and evolution of web page templates. In *International World Wide Web Conference*, pages 830–839, New York, NY, USA, 2005. ACM.

[81] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[82] Google. Web authoring statistics. `http://code.google.com/webstats/index.html`, retrieved 2010-11-05, Jan 2006.

[83] A. Gregorowicz and M. A. Kramer. Mining a Large–Scale Term–Concept Network from Wikipedia. Technical Report 06–1028, MITRE Corporation, MA, USA, Oct 2006.

[84] I. Gurevych. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005)*, pages 767–778. Springer, 2005.

[85] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03)*. IEEE Computer Society, 2003.

[86] S. Hassan and R. Mihalcea. Cross–lingual Semantic Relatedness Using Encyclopedic Knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201. Association for Computational Linguistics, 2009.

[87] F. Henri, B. Charlier, and F. Limpens. Understanding PLE as an Essential Component of the Learning Process. In *Proceedings of the 20th World Conference on Educational Multimédia Hypermedia & Telecommunications, EDMEDIA*, pages 3766–3770, 2008.

[88] I. Hickson et al. HTML5: A vocabulary and associated APIs for HTML and XHTML. `http://www.w3.org/TR/html5/`, retrieved 2010-11-15, Oct 2010.

[89] R. Hiemstra. *The International Encyclopedia of Education*, chapter Self–Directed Learning. Pergamon Press, Oxford, second edition, 1994.

[90] H. W. Hodgins. The Future of Learning Objects. In J. R. Lohmann and M. L. Corradini, editors, *Proceedings of e-Technologies in Engineering Education: Learning Outcomes Providing Future Possibilities*, number 1, pages 75–82, Davos, Switzerland, 2002. ECI Symposium Series.

[91] H. W. Hodgins, E. Duval, et al. IEEE 1484.12.1-2002, Draft Standard for Learning Object Metadata. `http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf`, retrieved 2010-10-13, 2002.

[92] S. Hoermann. *Wiederverwendung von digitalen Lernobjekten in einem auf Aggregation basierenden Autorenprozess*. PhD thesis, TU Darmstadt, Feb 2006.

[93] V. Hollink, J. Kamps, C. Monz, and M. De Rijke. Monolingual Document Retrieval for European Languages. *Information Retrieval*, 7:33–52, Jan 2004.

[94] T. Hug and N. Friesen. *Didactics of Microlearning. Concepts, Discourses and Examples*, chapter Outline of a Microlearning Agenda, pages 15–31. Waxmann Verlag, Barcelona, Spain, 2007.

[95] T. Hug and N. Friesen. Outline of a Microlearning Agenda. In *eLearning Papers*, number 16 in -, pages 1–13. elearningeuropa.info, Barcelona, Spain, Sep 2009.

[96] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, first edition, 2010.

[97] M. Jarmasz and S. Szpakowicz. Roget's Thesaurus and Semantic Similarity. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP*, 1:111, 2004.

[98] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 1997.

[99] F. Kaiser, H. Schwarz, and M. Jakob. Using Wikipedia-Based Conceptual Contexts to Calculate Document Similarity. *International Conference on the Digital Society*, 0:322–327, 2009.

[100] A. Kennedy and M. Shepherd. Automatic identification of home pages on the web. In *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4*, pages 99–103, Washington, DC, USA, 2005. IEEE Computer Society.

[101] B. Kessler, G. Numberg, and H. Schütze. Automatic detection of text genre. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, pages 32–38, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[102] D. Kinzler. Automatischer Aufbau eines multilingualen Thesaurus durch Extraktion semantischer und lexikalischer Relationen aus der Wikipedia. Master's thesis, Universität Leipzig, May 2008.

[103] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The Web as a graph: Measurements, models, and methods. In *Proceedings of the 5th annual international conference on Computing and Combinatorics*, pages 1–17. Springer-Verlag, 1999.

[104] M. S. Knowles. *Self-directed learning*. Cambridge Adult Education, 1975.

[105] P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit 2005*, volume 5, pages 79–86. European Association for Machine Translation, Sep 2005.

[106] P. Kolari, T. Finin, and A. Joshi. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*. Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Mar 2006. Also available as technical report TR-CS-05-13.

[107] A. Kolcz and W. Yih. Site–Independent Template–Block Detection. In *Knowledge Discovery in Databases: PKDD 2007*, volume 4702 of *Lecture Notes in Computer Science*, pages 152–163. Springer Berlin / Heidelberg, 2007.

[108] J. Kooken, T. Ley, and R. de Hoog. How Do People Learn at the Workplace? Investigating Four Workplace Learning Assumptions. In E. Duval, R. Klamma, and M. Wolpers, editors, *Proceedings of EC–TEL 2007*, volume 4753 of *LNCS*, pages 158–171. Springer Verlag Berlin Heidelberg, 2007.

[109] R. Koper. *Reusing Online Resources: A Sustainable Approach to eLearning*, chapter Combining reusable learning resources and services with pedagogical purposeful units of learning, pages 46–59. Kogan Page, 2003.

[110] S. Kopf, B. Guthier, H. Lemelson, and W. Effelsberg. Adaptation of web pages and images for mobile applications. In *Proceedings of IS&T/SPIE conference on multimedia on mobile devices*, volume 7256, pages 1–12, 2009.

[111] H. Krohne, B. Egloff, C. Kohlmann, and A. Tausch. Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica*, 42:139–156, 1996.

[112] J. J. L'Allier. Frame of reference: NETg's map to the products, their structure and core beliefs. Technical report, National Education Training Group, 1997.

[113] T. K. Landauer, P. W. Foltz, and D. Laham. An Introduction to Latent Semantic Analysis. *Discourse processes*, 25(2):259–284, 1998.

[114] G. P. Latham and E. A. Locke. Self–Regulation through Goal Setting. *Organizational Behavior and Human Decision Processes*, 50:212–247, 1991.

[115] J. Lave and E. Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK, 1991.

[116] L. Lehmann, C. Rensing, and R. Steinmetz. Effektive Nutzung von Medienressourcen aus dem betriebswirtschaftlichen Lebenszyklus eines Produkts zur Modellierung und Produktion von Trainingsinhalten am Beispiel der EXPLAIN-Plattform. In P. C. Peter Loos, Volker Zimmermann, editor, *Prozessorientiertes Authoring Management: Methoden, Werkzeuge und Anwendungsbeispiele für die Erstellung von Lerninhalten*, pages 183–200. Logos Verlag, Berlin, Jan 2008.

[117] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[118] R. Levering, M. Cutler, and L. Yu. Using Visual Features for Fine–Grained Genre Classification of Web Pages. In *HICSS '08: Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, pages 131–140, Washington, DC, USA, 2008. IEEE Computer Society.

[119] S.-H. Lin and J.-M. Ho. Discovering informative content blocks from Web documents. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 588–593, New York, NY, USA, 2002. ACM Press.

[120] S. N. Lindstädt, P. Scheir, R. Lokaiczyk, B. Kump, G. Beham, and V. Pammer. Knowledge Services for Work–integrated Learning. In *Proceedings of the European Conference on Technology Enhanced Learning (ECTEL)*, pages 234–244. Springer, Sep 2008.

[121] S. Lingam and S. Elbaum. Supporting End-Users in the Creation of Dependable Web Clips. In *Proceedings of the 16th international conference on World Wide Web*, pages 953–962, Banff, 2007. ACM.

[122] X. Liu, Y. Hui, W. Sun, and H. Liang. Towards Service Composition Based on Mashup. In *Congress on Services,*, pages 332–339, Salt Lake City, UT, 2007. IEEE.

[123] J. B. Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11(1 and 2):22–31, Mar 1968.

[124] G. Macgregor and E. McCulloch. Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool. *Library Review*, 55(5):291–300, 2006.

[125] D. Mann, P. Scholl, C. Rensing, and R. Steinmetz. Interaktive, Community–unterstützte Wissensnetze für persönliches Wissensmanagement in Ressourcen–basierten Lernkontexten. In C. Rensing and G. Rößling, editors, *Proceedings der Pre–Conference Workshops der 5. e–Learning Fachtagung Informatik - DeLFI 2007*, pages 35–42, Berlin, Sep 2007. Logos Verlag.

[126] E. Masie. Making Sense of Learning Specifications & Standards: A Decision Maker's Guide to their Adoption. `www.staffs.ac.uk/COSE/cosenew/s3_guide.pdf`, retrieved 2010-09-21, 2003. Second Edition.

[127] S. McCarron, M. Ishikawa, et al. XHTML 1.1 — Module–based XHTML — Second Edition. `http://www.w3.org/TR/xhtml11/`, retrieved 2011-01-21, Nov 2010.

[128] M. McHale. A comparison of WordNet and Roget's taxonomy for measuring semantic similarity. In *Workshop on Usage of WordNet in Natural Language Processing Systems (COLING-ACL 1998). Online Proceedings, available from http://xxx. lanl. gov/abs/cmp-lg/9809003*, 1998.

[129] N. Meder. *Web-Didaktik: Eine neue Didaktik webbasierten, vernetzten Lernens*. Bertelsmann, 2006.

[130] M. Meyer. *Modularization and Multi–Granularity Reuse of Learning Resources*. PhD thesis, Technische Universität Darmstadt, Sep 2008.

[131] M. Meyer, C. Rensing, and R. Steinmetz. Using Community–Generated Contents as a Substitute Corpus for Metadata Generation. *International Journal of Advanced Media and Communication*, 2(1):59–72, 2008.

[132] S. Meyer zu Eissen and B. Stein. Genre Classification of Web Pages — User Study and Feasibility Analysis. In *KI 2004: Advances in Artificial Intelligence*, volume 3238 of *LNCS*, pages 256–269. Springer Berlin / Heidelberg, 2004.

[133] P. Mika, M. Ciaramita, H. Zaragoza, and J. Atserias. Learning to tag and tagging to learn: A case study on Wikipedia. *IEEE Intelligent Systems*, 23(5):26–33, 2008.

[134] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30. AAAI Press, Chicago, USA, 2008.

[135] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA, 2008. ACM.

[136] T. A. Miloi. Ähnlichkeitsbasierte Vorverarbeitung von Webseiten. Master's thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2005.

[137] T. Mitamura, F. Lin, H. Shima, M. Wang, J. Ko, J. Betteridge, M. Bilotti, A. Schlaikjer, and E. Nyberg. JAVELIN III: Cross-lingual question answering from Japanese and Chinese documents. In *Proceedings of NTICIR-6 Workshop*, Tokyo, Japan, May 2007.

[138] T. M. Mitchell. *Machine Learning*. Computer Science Series — Artificial Intelligence. McGraw–Hill International Editions, 1997.

[139] J. Morris and G. Hirst. The Subjectivity of Lexical Cohesion in Text. In W. B. Croft, J. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 41–47. Springer Netherlands, 2006.

[140] M. Müller–Prove. *Good Tags — Bad Tags. Social Tagging in der Wissensorganisation*, chapter Modell und Anwendungsperspektive des Social Tagging, pages 16–22. Waxmann, 2008.

[141] J. L. Myers and A. Well. *Research design and statistical analysis*. Lawrence Erlbaum, second edition, 2003.

[142] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining japanese weblogs. In *WWW (Alternate Track Papers & Posters)*, pages 320–321. WWW, ACM, New York, NY, USA., May 2004.

[143] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers. Analyzing Entities and Topics in News Articles Using Statistical Topic Models. *Intelligence and Security Informatics*, 3975:93–104, 2006.

[144] P. Noufal. Metadata: Automatic generation and extraction. In *7th MANLIBNET Annual National Convention on Digital Libraries in Knowledge Management: Opportunities for Management Libraries*, pages 319–327, 2005.

[145] D. W. Oard and A. R. Diekema. Cross-Language Information Retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–256, 1998.

[146] T. O'Reilly. What is web 2.0 — design patterns and business models for the next generation of software. `http://oreilly.com/web2/archive/what-is-web-20.html`, retrieved 2010-11-05, Sep 2005.

[147] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, Jan 1998.

[148] S. Patwardhan, S. Banerjee, and T. Pedersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proceedings of Computational Linguistics and Intelligent Text Processing: 4th international conference, CICLing 2003*, pages 241–257, Mexico, 2003. Springer.

[149] J. C. Platt. *Fast training of support vector machines using sequential minimal optimization*, chapter 12, pages 185–208. MIT Press Cambridge, MA, USA, 1999.

[150] P. R. Polsani. Use and Abuse of Reusable Learning Objects. *Journal of Digital Information*, 3(4):1–10, Feb 2003.

[151] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, Jul 1980.

[152] M. Potthast, B. Stein, and M. Anderka. A Wikipedia–Based Multilingual Retrieval Model. In *Proceedings of the IR research, 30th European conference on Advances in Information Retrieval*, pages 522–530. Springer-Verlag, 2008.

[153] D. Rafiei, D. L. Moise, and D. Sun. Finding Syntactic Similarities Between XML Documents. In *Proceedings of the Conference on Database and Expert Systems Applications (DEXA'06)*, volume 0, pages 512–516, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[154] D. Raggett, A. Le Hors, I. Jacobs, et al. HTML 4.01 Specification. `http://www.w3.org/TR/html401/`, retrieved 2011-01-21, Dec 1999.

[155] G. C. Rakes. Using the Internet as a Tool in a Resource-Based Learning Environment. *Educational Technology*, 36(5):52–56, Sep 1996.

[156] L. Ramaswamy, I. Arun, L. Liu, and F. Douglis. Automatic detection of fragments in dynamically generated web pages. In *Proceedings of the 13th international conference on World Wide Web*, pages 443–454, New York, NY, USA, 2004. ACM.

[157] RDF Working Group. Resource Description Framework (RDF). `http://www.w3.org/standards/techs/rdf`, retrieved 2010-12-20, Feb 2004.

[158] G. Rehm. Towards Automatic Web Genre Identification. In *35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, volume 4, pages 101–111, Los Alamitos, CA, USA, 2002. IEEE Computer Society.

[159] C. Rensing, S. Bergsträßer, T. Hildebrandt, M. Meyer, B. Zimmermann, A. Faatz, L. Lehmann, and R. Steinmetz. Re-Use and Re-Authoring of Learning Resources - Definitions and Examples. Technical Report KOM-TR-2005-02, TU Darmstadt - Multimedia Communications Lab, Darmstadt, Nov 2005.

[160] C. Rensing, P. Scholl, D. Böhnstedt, and R. Steinmetz. Recommending and Finding Multimedia Resources in Knowledge Acquisition Based on Web Resources. In *Proceedings of 19th International Conference on Computer Communications and Networks*, pages 1–6. IEEE, IEEE eXpress Conference Publishing, Aug 2010.

[161] C. Rensing, B. Zimmermann, M. Meyer, L. Lehmann, and R. Steinmetz. Wiederverwendung von multimedialen Lernressourcen im Re-Purposing und Authoring by Aggregation. In P. Loos, V. Zimmermann, and P. Chikova, editors, *Prozessorientiertes Authoring Management: Methoden, Werkzeuge und Anwendungsbeispiele für die Erstellung von Lerninhalten*, pages 19–40. Logos Verlag, Berlin, Jan 2008.

[162] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI95*, 1995.

[163] D. Riboni. Feature Selection for Web Page Classification. In A. M. Tjoa, editor, *In EURASIA–ICT 2002 Proceedings of the Workshop*. Austrian Computer Society, 2002.

[164] H. Rubenstein and J. B. Goodenough. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[165] R. Saarso. HTML. `http://triin.net/2006/06/12/HTML`, retrieved 2010-11-05, Jun 2006.

[166] L. Salmerón, W. Kintsch, and E. Kintsch. Self-Regulation and Link Selection Strategies in Hypertext. *Discourse Processes*, 47(3):175–211, 2010.

[167] M. Santini. Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically–Motivated Features. Technical Report ITRI-05-02, University of Brighton, Apr 2005.

[168] M. Santini. Some Issues in Automatic Genre Classification of Web Pages. In *JADT 2006: 8es Journées Internationales d'Analyse statistique des Données Textuelles*, 2006.

[169] M. Santini. *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton, Jan 2007.

[170] A. Schiller, S. Teufel, C. Stöckert, and C. Thielen. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/stts_guide.pdf`, retrieved 2010-12-13, Nov 1995.

[171] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12, pages 44–49. Manchester, UK, 1994.

[172] S. Schmidt. Language–Independent Semantic Relatedness of Web Resources using Wikipedia as Reference Corpus. Master's thesis, Technische Univerität Darmstadt, Nov 2010.

[173] B. Schmitz and B. S. Wiese. New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology*, 31(1):64–96, 2006.

[174] P. Scholl, B. Benz, D. Böhnstedt, C. Rensing, B. Schmitz, and R. Steinmetz. Implementation and Evaluation of a Tool for Setting Goals in Self-Regulated Learning with Web Resources. In M. S. Ulrike Cress, Vania Dimitrova, editor, *Learning in the Synergy of Multiple Disciplines, EC-TEL 2009*, volume LNCS Vol 5794, Oct 2009.

[175] P. Scholl, B. Benz, D. Böhnstedt, C. Rensing, R. Steinmetz, and B. Schmitz. Einsatz und Evaluation eines Zielmanagement–Werkzeugs bei der selbstregulierten Internet–Recherche. In S. Seehusen, U. Lucke, and S. Fischer, editors, *DeLFI 2008: 6. e–Learning Fachtagung Informatik*, number P–132 in Lecture Notes in Informatics (LNI), pages 125–136, Köllen, Bonn, Sep 2008. Gesellschaft für Informatik, Springer.

[176] P. Scholl, D. Mann, C. Rensing, and R. Steinmetz. Support of Acquisition and Organization of Knowledge Artifacts in Informal Learning Contexts. In E. Distance and E.-L. Network, editors, *EDEN — Book of Abstracts*, page 16, Jun 2007.

[177] P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány. Cross–Language Retrieval with Wikipedia. *Advances in Multilingual and Multimodal Information Retrieval*, 5152:72–79, 2008.

[178] M. Shepherd, C. Watters, and A. Kennedy. Cybergenre: Automatic Identification of Home Pages on the Web. *Journal of Web Engineering*, 3(3&4):236–251, 2004.

[179] R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma. Learning important models for web page blocks based on layout and content analysis. *SIGKDD Explor. Newsl.*, 6(2):14–23, 2004.

[180] Y. Song, L. Zhang, and C. L. Giles. Automatic Tag Recommendation Algorithms for Social Recommender Systems. *ACM Transactions on the Web (TWEB)*, 5(1):4, 2011.

[181] P. Sorg and P. Cimiano. Cross–lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes for the CLEF 2008 Workshop*, 2008.

[182] P. Sorg and P. Cimiano. Enriching the Crosslingual Link Structure of Wikipedia — A Classification–Based Approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, pages 49–54, 2008.

[183] J. F. Sowa. Semantic Networks. In S. C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, volume 2, pages 1493–1511. John Wiley, New York, 1992.

[184] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. The JRC–Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *LREC 2006: Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006.

[185] J. Stenback, P. Le Hégaret, A. Le Hors, et al. Document Object Model HTML. `http://www.w3.org/TR/DOM-Level-2-HTML/html.html`, retrieved 2011-01-13, Jan 2003.

[186] M. Strube and S. P. Ponzetto. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1419ff, Menlo Park, CA; Cambridge, MA; London;, 2006. AAAI Press; MIT Press.

[187] S.-O. Tergan. *Knowledge and Information Visualization*, chapter Digital Concept Maps for Managing Knowledge and Information, pages 185–204. Springer Berlin / Heidelberg, 2005.

[188] N. Tintarev. Explanations of Recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 203–206. ACM, 2007.

[189] E. G. Toms and D. G. Campbell. Genre as Interface Metaphor: Exploiting Form and Function in Digital Environments. In *HICSS'99 — Proceedings of the 32nd Hawaii International Conference on System Sciences*, pages 341–357, Washington, DC, USA, Jan 1999. IEEE Computer Society.

[190] M. Tsukada, T. Washio, and H. Motoda. Automatic Web–Page Classification by Using Machine Learning Methods. In *WI '01: Proceedings of the First Asia–Pacific Conference on Web Intelligence: Research and Development*, pages 303–313, London, UK, 2001. Springer-Verlag.

[191] A. Ulbrich, P. Scheir, S. N. Lindstädt, and M. Görtz. A Context–Model for Supporting Work–Integrated Learning. In W. Nejdl and K. Tochtermann, editors, *Proceedings of EC–TEL 2006*, number 4227 in Lecture Notes of Computer Science, pages 525–530. EC-TEL, Springer-Verlag, Oct 2006.

[192] M. van Harmelen. Personal Learning Environments. In *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies*, pages 815–816. IEEE, 2006.

[193] K. Verbert and E. Duval. ALOCOM: a generic content model for learning objects. *International Journal on Digital Libraries*, 9:41–63, Jun 2008.

[194] L. S. Vygotsky. *Mind in Society. The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

[195] K.-P. Wild and U. Schiefele. Lernstrategien im Studium. Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 15:185–200, 1994.

[196] D. A. Wiley. Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. *The Instructional Use of Learning Objects*, 2830(435):1–35, 2000.

[197] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition, 2005.

[198] P. Xiang, X. Yang, , and Y. Shi. Effective page segmentation combining pattern analysis and visual separators for browsing on small screens. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 831–840, Washington, DC, USA, 2006. IEEE Computer Society.

[199] Z. Xu, I. King, and M. R. Lyu. Web Page Classification with Heterogeneous Data Fusion. In *Proceedings of the WWW Conference 2007*, pages 1171–1172. IW3C2, 2007.

[200] Y. Yang. An Evaluation of Statistical Approaches to Text Categorization. *Information retrieval*, 1(1):69–90, 1999.

[201] L. Yi, B. Liu, and X. Li. Eliminating noisy information in Web pages for data mining. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 296–305, New York, NY, USA, 2003. ACM Press.

[202] T. Zesch and I. Gurevych. Analysis of the wikipedia category graph for NLP applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pages 1–8, Rochester, New York, USA, 2007. Association for Computational Linguistics.

[203] T. Zesch and I. Gurevych. The More the Better? Assessing the Influence of Wikipedia's Growth on Semantic Relatedness Measures. In *Proceedings of the International Conference on Language Resources and Evaluation 2010*, Malta, May 2010. European Language Resources Association.

[204] T. Zesch, C. Müller, and I. Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 2008.

[205] B. Zimmermann. *Pattern–basierte Prozessbeschreibung und –unterstützung: Ein Werkzeug zur Unterstützung von Prozessen zur Anpassung von E–Learning–Materialien*. PhD thesis, Technische Universität Darmstadt, Dec 2008.

[206] G. K. Zipf. *The Psychobiology of Language*. Houghton-Mifflin, 1935.

[207] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1):1–10, Jul 2006.

List of Figures

List of Tables

List of Acronyms

API	Application Programming Interface	90
ASCII	American Standard Code for Information Interchange	122
BEP	Break Even Point	41
BPP	Blog Post Page	124
BSP	Blog Start Page	124
CBT	Computer Based Training	5
CL	Cross–Language	32
CSCL	Computer–supported Collaborative Learning	2
CSS	Cascading Stylesheets	5
DoC	Degree of Coherence	95
DOM	Document Object Model	90
ELWMS.KOM	E–Learning KnoWledge Management System	2
ESA	Explicit Semantic Analysis	2
FAQ	Frequently Asked Question	116
FSP	Forum Start Page	124
FTP	Forum Thread Page	124
HTML	HyperText Markup Language	5
HYRECA	Hybrid Recursive Segmentation Approach	90
IEEE	Institute of Electrical and Electronics Engineers	6
IR	Information Retrieval	26
LIGD	Language–Independent Web Genre Detection	3
LMS	Learning Management System	6
LO	Learning Object	2
LOM	Learning Object Metadata	6
LR	Learning Resource	15
LSA	Latent Semantic Analysis	31
LTSC	Learning Technology Standards Consortium	6
MAP	Mean Average Precision	41
MCL	Meta Cross–Language Link	64
MIME	Multipurpose Internet Mail Extensions	8
MP	Miscellaneous Page	125
NLP	Natural Language Processing	26
ODP	Open Directory Project	33

PAS	Pattern Analysis and visual Separators	96
PLE	Personal Learning Environment	13
POS	part–of–speech	57
RBL	Resource–Based Learning	1
RDF	Resource Description Framework	124
RIO	Reusable Information Object	9
RLO	Reusable Learning Object	9
RSS	Really Simple Syndication	13
SDL	Self–Directed Learning	13
SGML	Standard Generalized Markup Language	90
SRL	Self–Regulated Learning	3
SVM	Support Vector Machine	117
TEL	Technology Enhanced Learning	2
URL	Uniform Resource Locator	12
VIPS	Vision–Based Page Segmentation	95
VSM	Vector Space Model	30
W3C	World Wide Web Consortium	90
WBT	Web Based Training	5
WIL	Work–integrated Learning	15
WP	Wiki Page	124
WWW	World Wide Web	5
XESA	eXtended Explicit Semantic Analysis	2
XHTML	Extensible HyperText Markup Language	90
XML	Extensible Markup Language	90
XUL	XML User Interface Language	20

Appendix

A Appendix for Chapter Semantic Relatedness of Learning Resources

A.1 Corpora Samples

In this section, short samples for the datasets that were used for the different evaluations in chapter 3 are listed.

A.1.1 Relatedness of Term–Term Pairings

Rub65 (English)			Gur65 (German)			
term 1	term 2	Average	term 1	term 2	Average	STD
glass	tumbler	3.45	Glas	Becher	3.25	0.53
glass	jewel	1.78	Glas	Juwel	1.08	0.78
glass	magician	0.44	Glas	Zauberer	0.58	0.78
Examples for inconsistent translations						
shore	voyage	1.22	**Küste**	Reise	1.46	1.10
coast	**shore**	3.60	Küste	**Ufer**	3.67	0.48
rooster	cock	3.68	**Gockel**	Hahn	4.00	0.00
rooster	voyage	0.04	**Hahn**	Reise	0.00	0.00
Examples for incorrect translations						
sage	wizard	**2.46**	Fabel	Magier	**1.54**	1.17
oracle	sage	**2.61**	Orakel	Fabel	**1.25**	1.03
("sage" should be translated to "Weiser")						
cemetery	graveyard	**3.88**	Friedhof	Kirchhof	**3.00**	1.25
("graveyard" does not match "Kirchhof")						

Table A.1: Samples of English dataset Rub65 and corresponding German Gur65 dataset. The values refer to manual ratings of semantic relatedness. Rub65 does not provide standard deviations of values (STD). Note that the scale of both Rub65 and Gur65 range from 0 (not related) to 4 (identical). In the lower part of the table selected discrepancies between the datasets are shown, making the corpora unusable as parallel datasets for a cross–lingual evaluation.

term 1	term 2	Average	STD
Absage	ablehnen	3.50	0.534
Absage	Stellenanzeige	1.88	0.834
Affe	Gepäckkontrolle	0.13	0.353
Affe	Makake	4.00	0.000

Table A.2: Samples of German dataset Gur350. STD denotes standard deviations, the values refer to manual ratings of semantic relatedness. Note that the scale of Gur350 ranges from 0 (not related) to 4 (identical). In contrast to Gur65, this dataset contains more specific terms and verbs.

term 1 (en)	term 2 (en)	term 1 (de)	term 2 (de)	Average Relatedness
psychology	mind	Psychologie	Geist	7.69
five	month	fünf	Monat	3.38
planet	galaxy	Planet	Galaxie	8.11
vodka	gin	Wodka	Gin	8.46

Table A.3: Samples of parallel dataset `Schm280` for English and German. This dataset is based on the `wordsim353` dataset and has been translated by five raters. The ratings range from 0 (not related) to 10 (identical) and refer to manual ratings of semantic relatedness.

A.1.2 Relatedness of Query Term–Document Pairings

Query term	option 1	option 2	option 3	option 4
Bijou	Spitzbube	Spielkarte	**Schmuckstück**	Gaststätte
indiskret	**taktlos**	unaufdringlich	undurchschaubar	rücksichtslos
etepetete	begriffsstutzig	wichtigtuerisch	**zimperlich**	schüchtern
Compliance	Verwirrung	Vereinbarkeit	Vertrauen	**Therapietreue**

Table A.4: Samples of German dataset `RDWP984`. The correct answer is marked bold. The major challenge of this dataset is the inclusion of adjectives, rare and technical terminology.

A.1.3 Relatedness of Document–Document Pairings

Question ID	Answer ID	Text
8		Thema: Internet Slang. Was ist die Bedeutung folgender Worte oder Abkürzungen im Zusammenhang mit im Internet gebräuchlicher Sprache erläutern: kick, 1337, bot, imho, brb. Finden Sie pro Wort oder Abkürzung ein Snippet, das deren Bedeutung definiert.
	43	(I'll) be right back. <BRB> internet (Ich) bin gleich wieder da.
	44	Bot-Netz. Das kommt von robot und heißt soviel wie arbeiten. Im IT-Fachjargon ist mit Bot ein Programm gemeint, das ferngesteuert arbeitet.
	111	Entfernt einen User aus einem Channel, darf nur von einem Channel–Operator ausgeführt werden. Z.B.: Entfernt Fritzle mit dem Kommentar "und tschuess" aus dem Channel &Test: `KICK &Test Fritzle :und tschuess`.
	164	Ein Bot ist ein tendenziell eher simples, fleißiges "Arbeitswesen". Ungebräuchlich ist die Bezeichnung daher für quasi–selbständige Programme im Bereich der Künstlichen Intelligenz.

Table A.5: Sample of German `Gr282` dataset showing different answers to the question regarding common Internet slang. A peculiarity of this dataset is that it does not contain a one–to–one mapping of documents but contains documents that are self–similar in *semantic groups*.

ID	English sentence	German sentence
3	Provision for social, cultural and charitable causes must be included in Article 87 so that the future of these important institutions can also be permanently safeguarded.	In Artikel 87 muss der Bereich der Daseinsvorsorge für soziale, kulturelle und karitative Interessen mit aufgenommen werden, damit diese wichtigen Einrichtungen auch dauerhaft geschützt werden können.
4	Some people think that it is possible to create an internal market in Europe without agreeing on fundamental values and principles.	Einige sind der Meinung, dass man einen Binnenmarkt in Europa schaffen kann, ohne sich über grundlegende Werte und Prinzipien einig zu sein.
5	I too would like to welcome Mr Prodi's forceful and meaningful intervention.	Ich möchte meinerseits auch den klaren und substanziellen Redebeitrag von Präsident Prodi begrüßen.

Table A.6: Sample of English and German parallel sentences in `Europarl300` dataset. In an evaluation, one sentence is taken as a query and the respective parallel sentence should be found.

A.2 Questions in User Study for Semantic Corpus Gr282

The five participants of the user study to gather the data for the semantic corpus `Gr282` were asked to find snippets to answer the questions presented in table A.7.

ID	Group	Question	# Documents
1		Thema: aktuelle Politik. Finden Sie Snippets, die folgende Frage beantworten: Unter welchen Umständen dürfen gewählte Volksvertreter ein Dienstfahrzeug benutzen?	28
2		Finden Sie Snippets, die Meinungen oder Beschreibungen enthalten, ob der Begriff "dunkles Mittelalter" gerechtfertigt ist!	36
3		Finden Sie Snippets, die folgende Frage beantwortet: Was ist die FTAA?	31
4		Finden Sie Snippets, die die typischen Merkmale eines Abenteuerromans beschreiben.	27
5		Finden Sie Snippets, die folgende Frage beantwortet: Was ist Java?	
	5a	Objektorientierte Programmiersprache	24
	5b	Hauptinsel Indonesiens	3
	5c	Kaffeesorte	-
6		Finden Sie Snippets, die folgende Frage beantwortet: Was sind Puma, Jaguar, Panther, Tiger und Leopard?	
	6a	Betriebssysteme (Apple)	14
	6b	Tier (Schleichkatze)	13
	6c	Panzer	-
7		Finden Sie Snippets, die folgende Frage beantworten: Wie kann man selbst Bonbons herstellen?	33
8		Thema: Internet Slang. Was ist die Bedeutung folgender Worte oder Abkürzungen im Zusammenhang mit im Internet gebräuchlicher Sprache erläutern: *kick*, *1337*, *bot*, *imho*, *brb*. Finden Sie pro Wort oder Abkürzung ein Snippet, das deren Bedeutung definiert.	26
9		Finden Sie Snippets, die über die Entstehung Roms Auskunft geben. Sie müssen sich nicht ausschließlich auf historische Angaben beschränken.	
	9a	Historischer Verlauf	13
	9b	Mythologie	12
10		Finden Sie Snippets, die die Entwicklung des Menschen beschreiben. Sie müssen sich nicht ausschließlich auf wissenschaftlich belegte Angaben beschränken.	
	10a	Entwicklungspsychologie und Wachstum	-
	10b	Evolution	17
	10c	Schöpfungsgeschichte	5
	14		282

Table A.7: Questions asked in the user study for building the semantic corpus `Gr282`

A.3 Addendum for Filtering Strategy Evaluations

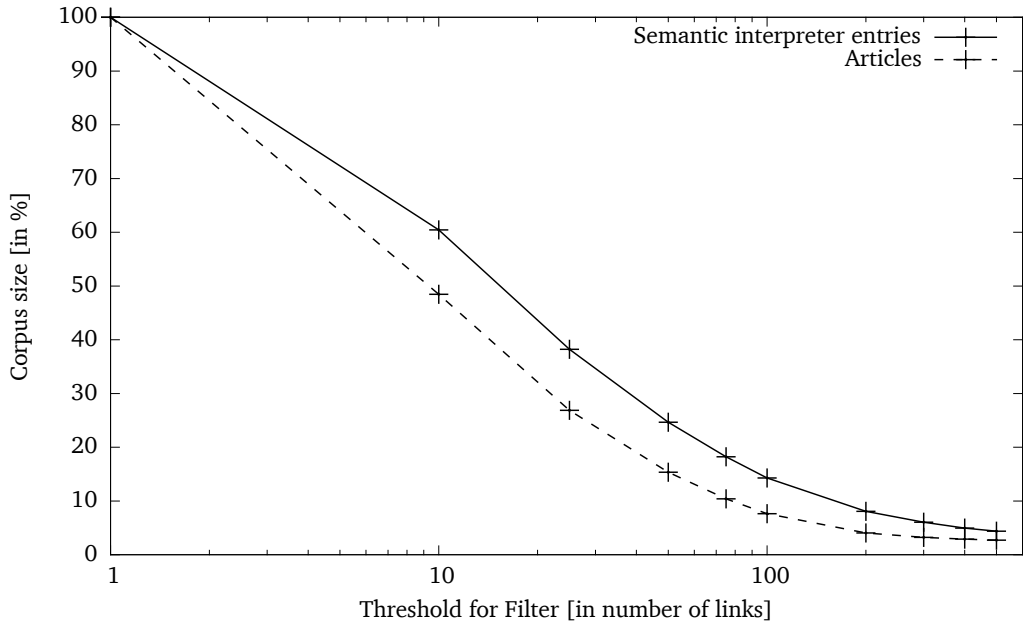


Figure A.1: Comparison of semantic interpreter size (in non–zero entries) and article count for inlink filtering strategy

	RDWP984					Gur65	Gur350
Link Threshold	Covered	Wrong	Correct	Global Accuracy	Local Accuracy	Correlation ρ	Correlation ρ
0	862	243	619	62.91%	71.81%	0.68	0.49
10	839	244	595	60.47%	70.92%	0.68	0.47
25	813	229	584	59.35%	71.83%	0.62	0.47
50	774	249	525	53.35%	67.83%	0.61	0.46
75	746	257	489	49.70%	65.55%	0.59	0.45
100	722	261	461	46.85%	63.85%	0.59	0.43
200	660	273	387	39.33%	58.64%	0.55	0.36
300	624	275	349	35.47%	55.93%	0.52	0.35
400	590	274	316	32.11%	53.56%	0.48	0.34
500	577	282	295	29.98%	51.13%	0.43	0.33

Table A.8: Performance of the inlink filter strategy for selected points of measure

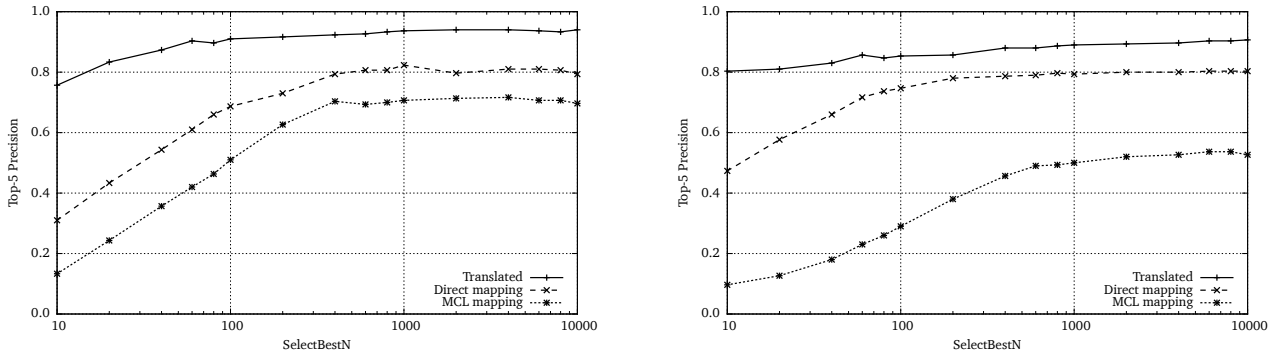
	RDWP984					Gur65	Gur350
Link Threshold	Covered	Wrong	Correct	Global Accuracy	Local Accuracy	Correlation ρ	Correlation ρ
0	862	243	619	62.91%	71.81%	0.68	0.49
25	834	262	572	58.13%	68.59%	0.65	0.43
45	800	280	520	52.85%	65.00%	0.58	0.40
50	796	283	513	52.13%	64.45%	0.55	0.40
75	764	302	462	46.95%	60.47%	0.54	0.39
100	733	308	425	43.19%	57.98%	0.48	0.38
200	648	317	331	33.64%	51.08%	0.48	0.31

Table A.9: Performance of the outlink filter strategy for selected points of measure

Link Threshold	Articles	Gr282 Break Even Point	Gr282 Mean Average Precision
0	973,227	0.6093	0.6238
10	471,504	0.6142	0.6284
25	261,569	0.6208	0.6338
50	149,370	0.6270	0.6427
75	101,082	0.6270	0.6451
100	74,270	0.6277	0.6521
200	39,406	0.6087	0.6331
300	31,452	0.5922	0.6066
400	28,078	0.5789	0.5937
500	26,436	0.5745	0.5873

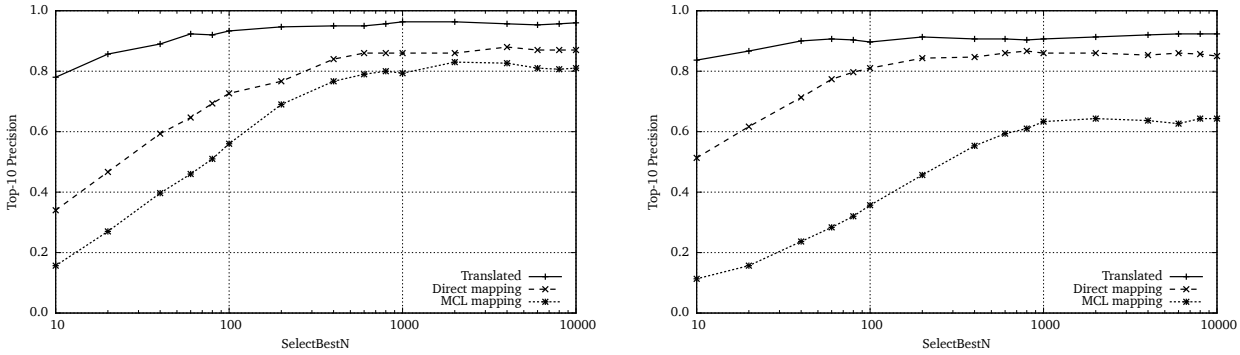
Table A.10: Impact of the inlink filter strategy on a document retrieval task performed on the Gr282 dataset

A.4 Addendum for Cross–Lingual ESA results



(a) English Target Language Space l_{en} **(b)** German Target Language Space l_{de}

Figure A.2: Top–5 Precision for Information Retrieval task using the `Europarl300` dataset with disambiguation page filter and the three different mapping strategies. The x–axis represents the considered number of most relevant concepts of interpretation vector i_{esa}. The translation result is computed using monolingual ESA in the given target language space.



(a) English Target Language Space l_{en} **(b)** German Target Language Space l_{de}

Figure A.3: Top–10 Precision for Information Retrieval task using the `Europarl300` dataset with disambiguation page filter and the three different mapping strategies. The x–axis represents the considered number of most relevant concepts of interpretation vector i_{esa}. The translation result is computed using monolingual ESA in the given target language space.

B Appendix for Chapter Granularity of Web Resources

B.1 Listing of HTML elements

This table contains all HTML4 elements that are covered by the structural analysis approaches described in chapters 4 (web resource segmentation) and 5 (web genre detection).

Element	Description	Type	Element	Description	Type
a	anchor	i	label	form field label text	i
abbr	abbreviated form (e.g. WWW, HTTP)	i	legend	fieldset legend	i
acronym	acronym (e.g. ELWMS.KOM)	i	li	list item	b
address	contact information on author	b	link	a media–independent link	i
applet	Java applet (deprecated, but still commonly used)	b	map	client–side image map	i
area	client-side image map area	i	menu	menu list	b
b	bold text style	i	meta	generic meta–information	i
big	large text style	i	noscript	alternate content container for non script–based rendering	i
blockquote	long quotation	b	object	generic embedded object	b
body	document body	b	ol	ordered list	b
br	forced line break	i	optgroup	option group	i
button	push button	i	option	selectable choice	i
caption	table caption	i	p	paragraph	b
cite	citation	i	param	named property value	i
code	computer code fragment	i	pre	preformatted text	b
col	table column	i	q	short inline quotation	i
colgroup	table column group	i	s	strike–through text style	i
dd	definition description	b	samp	sample program output, scripts, etc.	i
del	deleted text	i	script	script statements (e.g. *ECMAScript*)	i
dfn	instance definition	i	select	option selector	i
dir	directory list	b	small	small text style	i
div	generic language/style container	b	span	generic language/style container	i
dl	definition list	b	strike	strike-through text	i
dt	definition term	b	strong	strong emphasis	i
em	emphasis	i	style	styles (e.g. CSS)	i
fieldset	form control group	b	sub	subscript	i
font	local change to font	i	sup	superscript	i
form	interactive form	b	table	table with content structured in cells	b
frame	subwindow	b	tbody	table body	b
frameset	window subdivision	b	td	table data cell	b
hb-6	heading	b	textarea	multi–line text field	i
head	document head	i	tfoot	table footer	b
hr	horizontal rule	b	th	table header cell	b
html	document root element	i	thead	table header	b
i	italic text style	i	title	document title	i
iframe	inline subwindow	b	tr	table row	b
img	embedded image	i	tt	teletype or monospaced text style	i

Table B.1 continued on next page ...

... Table B.1 continued from previous page

Element	Description	Type	Element	Description	Type
input	form control	i	u	underlined text style	i
ins	inserted text	i	ul	unordered list	b
kbd	text to be entered by the user	i	var	instance of a variable or program argument	i

Table B.1: Listing of HTML4 elements. The type column denotes whether this respective tag is interpreted as a block level (b) or inline (i) element.

B.2 Web Pages contained in the Segmentation Corpus

The 10 biggest companies' homepages according to http://money.cnn.com/magazines/fortune/global500/2007/ were chosen for the company genre. All pages were crawled on the 9th of July 2008.

URL	$\overline{e_m}$	$\overline{e_s}$	$\overline{e_i}$	$\overline{e_b}$	$\overline{e_w}$	precision	recall
Blogs							
http://www.boingboing.net/	0.40	2.60	0.20	0.00	0.20	0.94	0.98
http://www.boingboing.net/2008/06/09/tsa-outlaws-flights.html	0.00	4.00	2.40	0.20	0.00	0.92	0.97
http://www.problogger.net/	0.00	3.00	0.40	0.00	0.00	0.92	0.99
http://www.problogger.net/archives/2008/06/06/top-10-plurk-users-statistics-whats-the-karma-algorithm/	0.20	0.40	0.60	0.00	0.00	0.98	0.98
http://www.techcrunch.com/	0.40	0.60	3.40	0.00	0.00	0.96	0.96
http://www.techcrunch.com/2008/06/09/the-3gps-iphone-arrives/	1.40	1.00	0.40	0.40	0.20	0.93	0.92
http://lifehacker.com/	2.00	3.40	2.00	1.20	0.00	0.89	0.91
http://lifehacker.com/395368/best-online-language-tools-for-word-nerds	0.60	3.40	0.20	0.00	0.20	0.95	0.99
http://www.engadget.com/	0.60	4.60	3.00	0.00	0.00	0.91	0.96
http://www.engadget.com/2008/06/09/the-lucky-22-countries-receiving-iphone-3g-on-july-11th	1.40	3.00	1.40	0.00	0.00	0.96	0.97
Web shops							
http://www.dell.com/content/products/category.aspx/vostronb?c=us&cs=04&l=en&s=bsd	0.40	1.40	1.40	1.40	0.40	0.87	0.90
http://www.ebay.com/	0.00	2.80	0.00	0.20	0.00	0.88	0.99
http://computers.ebay.com/_W0QQ_trksidZp3907Q2em21	1.20	2.20	0.20	0.40	0.00	0.93	0.95
http://computers.listings.ebay.com/Apple-Computers-Components_Apple-Laptops-Notebooks_W0QQfromZR4QQsacatZ111422QQsocmdZListingItemList	1.00	1.20	0.40	0.20	0.00	0.98	0.98
http://www.amazon.com/	1.00	2.00	4.20	0.00	8.60	0.82	0.83
http://www.amazon.com/books-used-books-textbooks/b/ref=sa_menu_bo0/105-7180827-4152405?%5Fencoding=UTF8&node=283155&pf_rd_m=ATVPDKIKX0DER&pf_rd_s=left-nav-1&pf_rd_r=1FAWTE861C0VWK6N21AH&pf_rd_t=101&pf_rd_p=328655101&pf_rd_i=507846	2.40	2.60	1.00	0.00	0.00	0.93	0.94
http://www.amazon.com/Movies-Entertainment-Books/b/ref=amb_link_1?ie=UTF8&node=4484&pf_rd_p=236786501&pf_rd_s=browse&pf_rd_t=101&pf_rd_i=86&pf_rd_m=ATVPDKIKX0DER&pf_rd_r=1YPRKT34DT0MDX7HBGS3	2.40	5.00	2.40	0.20	0.00	0.93	0.95
http://www.asos.com/Women/Dresses/Cat/pgecategory.aspx?cid=4168	0.60	0.40	0.00	0.80	0.20	0.96	0.95

Table B.2 continued on next page ...

... Table B.2 continued from previous page

URL	$\overline{e_m}$	$\overline{e_s}$	$\overline{e_i}$	$\overline{e_b}$	$\overline{e_w}$	precision	recall
Company homepages							
http://walmartstores.com/	0.40	0.40	0.80	0.00	0.00	0.91	0.91
http://www.exxonmobil.com/corporate/	0.20	1.00	0.00	0.00	0.00	0.93	0.99
http://www.shell.com/	1.20	0.60	0.20	0.00	0.00	0.92	0.87
http://www.bp.com/home.do?categoryId=1	0.20	2.40	0.40	0.00	0.00	0.80	0.95
http://www.daimler.com/	0.80	1.20	0.00	0.00	0.00	0.93	0.95
http://www.conocophillips.com/index.htm	1.60	0.40	0.60	0.60	0.00	0.95	0.92
http://www.total.com/en/home_page/	0.60	0.60	0.20	0.00	0.00	0.96	0.96
Miscellaneous							
http://www.apothekepforzheim.com/	1.40	0.00	0.00	0.00	0.00	1.00	0.68
http://tagesschau.de/showthread.php?p=755704&mode=linear.html	3.40	2.00	0.00	0.00	0.20	0.69	0.57
http://forum.tagesschau.de/forumdisplay.php?s=589c5141a3fe34342c2c1cec63d71b61&f=624	1.80	0.60	0.00	0.20	0.00	0.99	0.97
http://de.wikipedia.org/wiki/Nacktschnecke	1.80	2.00	3.80	0.20	0.00	0.93	0.93
http://moinmoin.wikiwikiweb.de/MoinMoinExtensions	2.20	2.80	0.40	0.20	0.00	0.69	0.73
http://www.pythonware.com/daily/	2.20	0.20	0.20	0.20	0.00	0.97	0.88
http://www.kom.tu-darmstadt.de/en/research/research-areas/	1.00	0.00	0.20	0.00	0.00	0.99	0.92
http://www.elearningpost.com/	0.40	0.00	0.00	0.00	0.00	1.00	0.99
http://www.microsoft.com/de/de/default.aspx	2.20	0.40	0.20	1.20	0.00	0.89	0.81
http://kunstmuehle.de/	0.40	0.60	0.20	0.00	0.20	0.83	0.86
News sites							
http://www.heise.de	1.20	0.60	0.60	0.40	0.00	0.97	0.96
http://www.heise.de/newsticker/Reframe-Seltenes-Filmmaterial-fuer-alle--/meldung/10921	0.60	0.40	0.40	0.00	0.00	0.96	0.95
http://www.spiegel.de	0.80	0.20	1.00	0.00	0.00	0.98	0.97
http://www.spiegel.de/spiegel/0,1518,558600,00.html	0.80	2.20	1.00	0.00	0.00	0.91	0.95
http://www.welt.de/	1.60	2.40	3.20	0.20	0.00	0.94	0.95
http://www.welt.de/politik/article2083198/Erding-Gipfel_zementiert_tiefen_Riss_in_der_Union.html	0.60	5.20	0.20	0.00	0.20	0.91	0.98
http://www.rp-online.de/public/article/sport/fussball/nationalelf/euro_2008/576970/Was-muss-uns-noch-Sorgen-machen.html	0.40	1.20	1.40	0.00	0.00	0.94	0.96
http://times.com/	3.20	1.40	1.00	0.40	0.00	0.96	0.94
http://www.nytimes.com/2008/06/10/us/09cnd-weather.html?_r=1&hp&oref=slogin	0.00	0.80	1.40	0.00	0.00	0.93	0.95

Table B.2: Listing of URLs of all web pages used for the evaluation. The different $\overline{e_x}$ denote the average of errors that were counted by the participants, with x being [m]issing, [s]uperfluous, [i]ncomplete, too [b]ig and completely [w]rong. Precision and recall are calculated based on these error classes (cf. chapter 4.5.2).

Pages that could not be rendered by Cobra:

- http://www.gm.com
- http://www.toyota.co.jp
- http://www.chevron.com
- http://www.focus.de

C Appendix for Chapter Web Genres as Metadata

C.1 Feature Details

In this section, details of the presented features that are used for Web Genre Detection (cf. chapter 5) are listed.

Facet	Contained HTML tags
Layout facet	`br`, `dl`, `dir`, `div`, `hr`, `ol`, `p`, `pre`, `table`, `ul`
Typographic facet	`abbr`, `acronym`, `address`, `b`, `big`, `blockquote`, `caption`, `center`, `cite`, `em`, `font`, `h1-6`, `i`, `img`, `q`, `s`, `small`, `strike`, `strong`, `style`, `sub`, `sup`, `tt`, `u`
Functionality facet	`applet`, `bgsound`, `button`, `embed`, `fieldset`, `form`, `input`, `legend`, `noscript`, `object`, `option`, `optgroup`, `param`, `script`, `select`, `textarea`, `var`, `mailto:`–links

Table C.1: Compilation of HTML tags for each aggregating facet feature (cf. [167]).

Average Merit	Average Rank	Feature
0.683 +- 0.014	1 +- 0	tag_freq_td
0.629 +- 0.01	2 +- 0	url_depth*
0.603 +- 0.014	3 +- 0	tag_freq_tr
0.58 +- 0.013	4 +- 0	tag_freq_table
0.481 +- 0.015	5.6 +- 0.66	tag_freq_h2
0.476 +- 0.021	6.3 +- 1.49	link_ratio_anchor*
0.461 +- 0.019	7.5 +- 1.36	tag_freq_p
0.455 +- 0.017	8.2 +- 1.89	tag_freq_input
0.437 +- 0.016	9.9 +- 1.81	pattern_median*
0.434 +- 0.011	10.4 +- 1.62	tag_freq_tbody
0.427 +- 0.026	11.6 +- 3.83	tag_freq_meta
0.425 +- 0.012	11.7 +- 1.42	tag_freq_strong
0.422 +- 0.013	12.1 +- 1.7	tag_freq_h3
0.41 +- 0.015	13.8 +- 1.99	tag_freq_img
0.405 +- 0.005	14.3 +- 1.27	tag_freq_thead
0.387 +- 0.011	16.7 +- 1.35	facets_func
0.389 +- 0.009	16.8 +- 1.33	tag_freq_li
0.386 +- 0.016	16.9 +- 2.12	tag_freq_ul
0.37 +- 0.015	18.7 +- 1.55	pattern_ratio_text*
0.356 +- 0.017	20.9 +- 1.97	tag_freq_option
0.351 +- 0.01	21.5 +- 1.02	tag_freq_h1
0.344 +- 0.01	22 +- 1.18	feed*
0.332 +- 0.028	23.9 +- 3.27	tag_freq_optgroup
0.328 +- 0.008	24.2 +- 1.47	link_ratio_www*
0.326 +- 0.006	24.2 +- 0.87	tag_freq_form
0.313 +- 0.008	26.6 +- 1.2	punctuation_gt
0.309 +- 0.017	27.4 +- 3.01	tag_freq_label
0.307 +- 0.022	27.7 +- 3.1	facets_typo
0.3 +- 0.01	28.3 +- 1.42	pattern_ratio*

Table C.2 continued on next page ...

... Table C.2 continued from previous page

Average Merit	Average Rank	Feature
0.296 +- 0.011	29.9 +- 2.55	facets_int_nav
0.285 +- 0.009	32.2 +- 1.99	tag_freq_link
0.285 +- 0.014	32.4 +- 2.46	pattern_nr*
0.276 +- 0.016	33.2 +- 3.43	link_ratio_site*
0.279 +- 0.01	33.4 +- 2.11	pattern_size*
0.269 +- 0.009	34.9 +- 1.14	punctuation_cm
0.264 +- 0.021	36.1 +- 4.35	tag_freq_div
0.254 +- 0.012	38.1 +- 2.91	punctuation_sh
0.251 +- 0.009	38.6 +- 2.54	pattern_start*
0.25 +- 0.006	38.8 +- 1.78	punctuation_co
0.243 +- 0.009	40.6 +- 2.8	tag_freq_h5
0.241 +- 0.013	41.2 +- 2.79	tag_freq_select
0.242 +- 0.011	41.3 +- 2.9	class_freq
0.24 +- 0.005	41.7 +- 1.9	css_rules*
0.232 +- 0.018	43.6 +- 3.98	text_ratio*
0.227 +- 0.007	44.6 +- 1.5	link_ratio_page*
0.223 +- 0.017	45.8 +- 3.19	css_bytes
0.222 +- 0.009	46.1 +- 1.81	block_tag_ratio*
0.207 +- 0.006	48.7 +- 1	tag_freq_hr
0.202 +- 0.015	49.7 +- 3.16	tag_freq_br
0.197 +- 0.008	50.5 +- 1.91	tag_freq_textarea

Table C.2: List of the top 50 features ranked by Information Gain. The features marked with * are novel and are proposed in this thesis.

C.2 List of Web Genres contained in Class *Miscellaneous Pages*

In this section, all genres that were taken as a starting point to assemble web resources for the MP class are listed. The web genres have been assembled from related work[1]. Having defined the genres of interest, at least ten web resources in different languages were manually searched for using Google[2]. A list of all URLs of the genre corpus can be downloaded from `http://pcscholl.de/thesis/genre_detection_urls.zip`.

E–shop / Corporate

- Commercial homepage, Promotional page
- Corporate homepage
- Product list / specification / document
- Advertisement
- Legal
- Portal, Movie / Restaurant critic

[1] A comprehensive listing can be found on `http://www.webgenrewiki.org/index.php5/Genre_Classes_List`, *retrieved 2011-02-22*

[2] `http://google.com`, *retrieved 2011-01-19*

News

- Reportage, Feature
- Editorial, Miscellaneous article
- Newswire
- Radio / Television news
- Technology news

Academic

- Academic / University homepage
- Call for papers, Conference / Workshop homepage
- Syllabus
- Research Report, Article, Technical / White paper
- Course / Department / Faculty / Project / Student homepage

Information

- Help, FAQ, How–to, Tutorial
- Manual
- Link collection, Sitemap
- Best practice, Tech–note, Presentation

Assorted

- Official homepage (e.g. city council)
- Content / Media sites
- Error Message (e.g. `404 - Page not found`)
- Index, Web directory, Download, Image collection
- Search engine
- Job description, Resume, C.V.
- Community
- Fiction, Poetry
- Entertainment (e.g. game sites)

C.3 Ranking of Delicious Tags

Rank	Tag name	Number of occurrences	Ratio
1	`design`	851,909	13.19‰
2	`software`	665,123	10.30‰
3	`blog`	625,938	9.69‰
4	`tools`	584,742	9.05‰
5	`programming`	567,858	8.79‰

Table C.3 continued on next page ...

... Table C.3 continued from previous page

Rank	Tag name	Number of occurrences	Ratio
6	reference	530,234	8.21‰
7	web	524,602	8.12‰
8	music	523,321	8.10‰
9	web2.0	461,420	7.14‰
10	video	457,336	7.08‰
11	webdesign	436,694	6.76‰
12	art	419,100	6.49‰
13	howto	372,679	5.77‰
14	linux	368,329	5.70‰
15	css	366,232	5.67‰
16	tutorial	329,085	5.09‰
17	photography	309,456	4.79‰
18	javascript	309,343	4.79‰
19	free	306,751	4.75‰
20	business	290,601	4.50‰
21	google	283,098	4.38‰
22	news	274,039	4.24‰
23	tips	269,249	4.17‰
24	development	259,916	4.02‰
25	blogs	257,823	3.99‰
26	politics	255,526	3.96‰
27	mac	251,403	3.89‰
28	opensource	241,722	3.74‰
29	windows	239,799	3.71‰
30	ajax	238,057	3.68‰
31	flash	235,063	3.64‰
32	security	229,259	3.55‰
33	shopping	227,870	3.53‰
34	education	227,707	3.52‰
35	technology	224,360	3.47‰
36	science	223,632	3.46‰
37	java	213,190	3.30‰
38	search	212,745	3.29‰
39	books	210,426	3.26‰
40	inspiration	209,937	3.25‰
41	internet	208,593	3.23‰
42	fun	198,225	3.07‰
43	games	198,052	3.07‰
44	humor	189,914	2.94‰
45	funny	182,039	2.82‰
46	travel	179,385	2.78‰
47	research	174,504	2.70‰
48	history	172,545	2.67‰
49	food	168,175	2.60‰
50	cool	166,176	2.57‰

Table C.3: Top 50 of ranked Delicious tags (cf. chapter 5.1)

D List of Own Publications

First Author

1. Philipp Scholl, Doreen Böhnstedt, Christoph Rensing, Ralf Steinmetz: *Adaptivity and Adaptability in Personal and Community Knowledge Networks*. In: Andreas Kaminski, Max Mühlhäuser, Werner Sesink, Jürgen Steimle (Eds.): Interdisciplinary Approaches to Technology Enhanced Learning. Interdisziplinäre Zugänge zu technologie–gestütztem Lernen. Münster: Waxmann, 2011 [in preparation].
2. Philipp Scholl, Andreas Kaminski: *Adaptivität — ein vierwertiger Begriff*. In: Andreas Kaminski, Max Mühlhäuser, Werner Sesink, Jürgen Steimle (Eds.): Interdisciplinary Approaches to Technology Enhanced Learning. Interdisziplinäre Zugänge zu technologie–gestütztem Lernen. Münster: Waxmann, 2011 [in preparation].
3. Philipp Scholl, Doreen Böhnstedt, Renato Domínguez García, Christoph Rensing, Ralf Steinmetz: *Extended Explicit Semantic Analysis for Calculating Semantic Relatedness of Web Resources*. In: Martin Wolpers, Paul A. Kirschner, Maren Scheffel, Stefanie Lindstädt, Vania Dimitrova: Sustaining TEL: From Innovation to Learning and Practice Proceedings of EC–TEL 2010, vol. Lecture Notes in Computer Science 6383, pp. 324–339, Springer Verlag, September 2010. ISBN 978-3-642-16019-6.
4. Philipp Scholl, Bastian F. Benz, Doreen Böhnstedt, Christoph Rensing, Bernhard Schmitz, Ralf Steinmetz: *Implementation and Evaluation of a Tool for Setting Goals in Self-Regulated Learning with Web Resources*. In: Ulrike Cress, Vania Dimitrova, Marcus Specht: Learning in the Synergy of Multiple Disciplines, EC–TEL 2009, LNCS Vol 5794, pp. 521-534, Springer-Verlag Berlin Heidelberg, October 2009. ISBN 3-642-04635-5.
5. Philipp Scholl, Doreen Böhnstedt, Renato Domínguez García, Christoph Rensing, Ralf Steinmetz: *Anwendungen und Nutzen der Automatischen Erkennung von Web-Genres in persönlichen und Community- Wissensnetzen*. In: Andreas Schwill, Nicolas Apostopoulos: Lernen im digitalen Zeitalter - Workshop-Band - Dokumentation der Pre-Conference zur DeLFI 2009, pp. 37–44, Logos, September 2009. ISBN 978-3-83252-273-5.
6. Philipp Scholl, Renato Domínguez García, Doreen Böhnstedt, Christoph Rensing, Ralf Steinmetz: *Towards Language–Independent Web Genre Detection*. In: ACM: WWW '09: Proceedings of the 18th international conference on World Wide Web, pp. 1157–1158, April 2009. ISBN 978-1-60558-487-4.
7. Philipp Scholl, Bastian Benz, Doreen Böhnstedt, Christoph Rensing, Ralf Steinmetz, Bernhard Schmitz: *Einsatz und Evaluation eines Zielmanagement–Werkzeugs bei der selbstregulierten Internet-Recherche*. In: Silke Seehusen, Ulrike Lucke, Stefan Fischer: DeLFI 2008: 6. e–Learning Fachtagung Informatik, no. P–132, pp. 125–136, Lecture Notes in Informatics (LNI), September 2008. ISBN 978-3-88579-226-0.
8. Philipp Scholl, Bastian Benz, Doreen Mann, Christoph Rensing, Ralf Steinmetz, Bernhard Schmitz: *Scaffolding von selbstreguliertem Lernen in einer Rechercheumgebung für internetbasierte Ressourcen*. In: Christoph Rensing, Guido Rößling: Proceedings der Pre–Conference Workshops der 5. e–Learning Fachtagung Informatik — DeLFI 2007, pp. 43–50, Logos Verlag, September 2007. ISBN 978-3-8325-1674-1.

9. Philipp Scholl, Doreen Mann, Christoph Rensing, Ralf Steinmetz: *Support of Acquisition and Organization of Knowledge Artifacts in Informal Learning Contexts*. In: European Distance and E–Learning Network: EDEN — Book of Abstracts, pp. 16–17, June 2007. ISBN 978-963-06-2655-2.

Co–Author

1. Doreen Böhnstedt, Philipp Scholl, Christoph Rensing, Ralf Steinmetz: *Enhancing an Environment for Knowledge Acquisition based on Web Resources by Automatic Tag Type Identification*. In: Michael E. Auer, Jeanne Schreurs: Proceedings of International Conference on Computer-aided Learning 2010 (ICL 2010), pp. 380–389, kassel university press, September 2010. ISBN 978-3-89958-541-4.
2. Christoph Rensing, Philipp Scholl, Doreen Böhnstedt, Ralf Steinmetz: *Recommending and Finding Multimedia Resources in Knowledge Acquisition Based on Web Resources*. In: Proceedings of 19th International Conference on Computer Communications and Networks, pp. 1–6, IEEE eXpress Conference Publishing, August 2010. ISBN 978-1-42447-114-0.
3. Renato Domínguez García, Alexandru Berlea, Philipp Scholl, Doreen Böhnstedt, Christoph Rensing, Ralf Steinmetz: *Improving Topic Exploration in the Blogosphere by Detecting Relevant Segments*. In: Klaus Tochtermann, Hermann Maurer: Proceedings of 9th International Conference on Knowledge Management and Knowledge Technologies (I–KNOW'09), pp. 177–188, Verlag der Technischen Universität Graz, September 2009. ISBN 978-3-85125-060-2.
4. Doreen Böhnstedt, Philipp Scholl, Christoph Rensing, Ralf Steinmetz: *Modeling Personal Knowledge Networks to Support Resource Based Learning*. In: Klaus Tochtermann, Hermann Maurer: Proceedings of 9th International Conference on Knowledge Management and Knowledge Technologies (I–KNOW'09), pp. 309–316, Verlag der Technischen Universität Graz, Austria, Universiti Malaysia Sarawak, Malaysia, and Know-Center, Austria, September 2009. ISBN 978-3-85125-060-2.
5. Renato Domínguez García, Doreen Böhnstedt, Philipp Scholl, Christoph Rensing, Ralf Steinmetz: *Von Tags zu semantischen Netzen — Einsatz im Ressourcen-basierten Lernen*. In: Andreas Schwill, Nicolas Apostopoulos: Lernen im digitalen Zeitalter — Workshop-Band — Dokumentation der Pre–Conference zur DeLFI 2009, pp. 29–36, Logos, September 2009. ISBN 978-3-83252-273-5.
6. Doreen Böhnstedt, Philipp Scholl, Christoph Rensing, Ralf Steinmetz: *Collaborative Semantic Tagging of Web Resources on the Basis of Individual Knowledge Networks*. In: Houben, G.-J.; McCalla, G.; Pianesi, F.; Zancanaro, M.: Proceedings of First and Seventeenth International Conference on User Modeling, Adaptation, and Personalization UMAP 2009, Lecture Notes in Computer Science vol. 5535, pp. 379–384, Springer-Verlag Berlin Heidelberg 2009, June 2009. ISBN 978-3-64202-246-3.
7. Simone Scherer, Philipp Scholl, Jan Hansen, Ralf Steinmetz: *eDemocracy — Weblogscreening zur frühzeitigen Erkennung von öffentlichen Meinungen*, ISPRAT e.V. Hamburg, March 2009.
8. Renato Domínguez García, Philipp Scholl, Doreen Böhnstedt, Christoph Rensing, Ralf Steinmetz: *Towards To an Automatic Web Genre Classification*. Technical Report no. TR-2008-10, November 2008.
9. Doreen Böhnstedt, Philipp Scholl, Bastian Benz, Christoph Rensing, Ralf Steinmetz, Bernhard Schmitz: *Einsatz persönlicher Wissensnetze im Ressourcen–basierten Lernen*. In: Silke Seehusen, Ulrike Lucke, Stefan Fischer: DeLFI 2008: 6. e–Learning Fachtagung Informatik, no. P-132, pp. 113–124, Lecture Notes in Informatics (LNI), September 2008. ISBN 978-3-88579-226-0.
10. Doreen Böhnstedt, Philipp Scholl, Christoph Rensing, Ralf Steinmetz: *ELWMS.KOM — Typisiertes Tagging in persönlichen Wissensnetzen*. In: Ulrike Lucke, Martin Christof Kindsmüller, Stefan Fischer, Michael Herczeg, Silke Seehusen: Workshop Proceedings der Tagungen Mensch & Computer 2008,

DeLFI 2008 und Cognitive Design 2008, pp. 330–331, Logos Verlag, September 2008. ISBN 978-3-8325-2007-6.

11. Doreen Mann, Philipp Scholl, Christoph Rensing, Ralf Steinmetz: *Interaktive, Community-unterstützte Wissensnetze für persönliches Wissensmanagement in Ressourcen–basierten Lernkontexten*. In: Christoph Rensing, Guido Rößling: Proceedings der Pre-Conference Workshops der 5. e–Learning Fachtagung Informatik — DeLFI 2007, pp. 35–42, Logos Verlag, September 2007. ISBN 978-3-8325-1674-1.

E List of Supervised Student's Theses

Master Theses and Diploma Theses

1. Vladislav Shakhmatov: *Automatic Keyphrase Extraction: An Approach for Filtering Search Results*, Master Thesis, Technische Universität Darmstadt, January 2011
2. Tobias Waller: *Semantic Analysis of Web Documents using Wikipedia*, Master Thesis, November 2010
3. Sebastian Schmidt: *Language–Independent Semantic Relatedness of Web Resources using Wikipedia as Reference Corpus*, Master Thesis, Technische Universität Darmstadt, November 2010
4. Johannes Grimm: *Berechnung semantischer Ähnlichkeit kleiner Textfragmente mittels Wikipedia*, Master Thesis, Technische Universität Darmstadt, October 2009
5. Anselm Föhr: *Extraction of Structurally Coherent Segments from Web Pages Using a Hybrid Recursive Approach*, Diploma Thesis, Technische Universität Darmstadt, August 2008
6. Dessislava Karaboneva: *Kombination von Methoden zum Management und zur Visualisierung von Wissen und Informationen*, Master Thesis, Technische Universität Darmstadt, February 2008

Student Research Projects

1. Ken Knoll: *Semantische Anreicherung von typisierten Tags durch Aggregation heterogener Datenquellen*, Student Research Project, Technische Universität Darmstadt, July 2010
2. Christopher Chard: *Konzept und Umsetzung einer interaktiven Visualisierung von Persönlichen und Community–Wissensnetzen*, Student Research Project, Technische Universität Darmstadt, July 2010
3. Artur Kuhn: *Identifikation von Online–Communities*, Student Research Project, Technische Universität Darmstadt, June 2008
4. Mohamad Kain: *Entwicklung eines Werkzeugs zum Zielmanagement in selbstbestimmten Lernprozessen*, Student Research Project, Technische Universität Darmstadt, February 2008

F Curriculum Vitae

Persönliche Daten

Name	Philipp Claudius Friedrich–Eugen Scholl
Geburtsdatum	21. Dezember 1979
Geburtsort	Stuttgart
Nationalität	deutsch
Familienstand	ledig, 1 Kind

Akademischer Werdegang

seit 11/2009	Technische Universität Darmstadt, Fachgebiet Multimedia Kommunikation — Wissenschaftlicher Mitarbeiter
11/2006–11/2009	Stipendiat des interdisziplinären Graduiertenkollegs *Qualitätsverbesserung im E–Learning durch rückgekoppelte Prozesse* der Technischen Universität Darmstadt
10/2000–09/2006	Studium der Medieninformatik an der Universität Ulm mit Abschluss *Diplom–Informatiker*
02/2003–02/2004	Auslandsstudium in Melbourne, Australien an der Monash University mit Abschluss *Bachelor of Multimedia Computing* (BMC)
08/1990–06/1999	Albert–Schweitzer–Gymnasium Leonberg mit Abschluss *Abitur*

Berufliche Tätigkeiten

seit 11/2009	Technische Universität Darmstadt, Fachgebiet Multimedia Kommunikation — Wissenschaftlicher Mitarbeiter
04/2007–06/2009	Technische Universität Darmstadt, Fachgebiet Multimedia Kommunikation — Wissenschaftliche Hilfskraft
04/2005–06/2005	Praktikum bei der *proKoncept GmbH*, Wolfratshausen
10/2004–02/2005	Universität Ulm, Entwicklung der Lernplattform *ShiFu* — Wissenschaftliche Hilfskraft
08/2004–10/2004	Praktikum beim *FWU Institut für Film und Bild in Wissenschaft und Unterricht*, München
04/2004–09/2004	Universität Ulm, Tutor für Vorlesung *Einführung in die Medienpädagogik, –psychologie und –didaktik* — Wissenschaftliche Hilfskraft
10/2002–02/2003	Universität Ulm, Tutor für Vorlesung *Praktische Informatik I* — Wissenschaftliche Hilfskraft
09/1999–08/2000	Zivildienst im *Diakonissenkrankenhaus Stuttgart* in der Onkologie

G Erklärung laut §9 der Promotionsordnung

Ich versichere hiermit, dass ich die vorliegende Dissertation allein und nur unter Verwendung der angegebenen Literatur verfasst habe. Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 2011

Dipl.–Inf. Philipp Claudius
Friedrich–Eugen Scholl